Lecture Notes in Electrical Engineering 619

Zahriladha Zakaria
Rabiah Ahmad *Editors*

Advances in Electronics Engineering

Proceedings of the ICCEE 2019, Kuala Lumpur, Malaysia

Springer

Lecture Notes in Electrical Engineering

Volume 619

The book series *Lecture Notes in Electrical Engineering* (LNEE) publishes the latest developments in Electrical Engineering—quickly, informally and in high quality. While original research reported in proceedings and monographs has traditionally formed the core of LNEE, we also encourage authors to submit books devoted to supporting student education and professional training in the various fields and applications areas of electrical engineering. The series cover classical and emerging topics concerning:

- Communication Engineering, Information Theory and Networks
- Electronics Engineering and Microelectronics
- Signal, Image and Speech Processing
- Wireless and Mobile Communication
- Circuits and Systems
- Energy Systems, Power Electronics and Electrical Machines
- Electro-optical Engineering
- Instrumentation Engineering
- Avionics Engineering
- Control Systems
- Internet-of-Things and Cybersecurity
- Biomedical Devices, MEMS and NEMS

For general information about this book series, comments or suggestions, please contact leontina.dicecco@springer.com.

To submit a proposal or request further information, please contact the Publishing Editor in your country:

China

Jasmine Dou, Associate Editor (jasmine.dou@springer.com)

India, Japan, Rest of Asia

Swati Meherishi, Executive Editor (Swati.Meherishi@springer.com)

Southeast Asia, Australia, New Zealand

Ramesh Nath Premnath, Editor (ramesh.premnath@springernature.com)

USA, Canada:

Michael Luby, Senior Editor (michael.luby@springer.com)

All other Countries:

Leontina Di Cecco, Senior Editor (leontina.dicecco@springer.com)

**** Indexing: The books of this series are submitted to ISI Proceedings, EI-Compendex, SCOPUS, MetaPress, Web of Science and Springerlink ****

More information about this series at http://www.springer.com/series/7818

Zahriladha Zakaria · Rabiah Ahmad
Editors

Advances in Electronics Engineering

Proceedings of the ICCEE 2019, Kuala Lumpur, Malaysia

Springer

Editors

Zahriladha Zakaria
Melaka, Malaysia

Rabiah Ahmad
Melaka, Malaysia

ISSN 1876-1100 ISSN 1876-1119 (electronic)
Lecture Notes in Electrical Engineering
ISBN 978-981-15-1288-9 ISBN 978-981-15-1289-6 (eBook)
https://doi.org/10.1007/978-981-15-1289-6

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

Foreword

Selamat Datang! It is our great pleasure to welcome you to the International Conference on Computer Science, Electrical and Electronic Engineering (ICCEE 2019) at Seri Pacific Hotel Kuala Lumpur, Malaysia, from 29 to 30 April 2019.

ICCEE 2019 is a comprehensive conference covering various topics of computer science, electrical and electronic engineering. We believe that by bringing together these related fields, it helps to create a unique opportunity for collaboration between industry and academic researchers. In addition, this conference will be a chance for discussing the existing challenging problems or those that emerge in industry. To all participants, we hope you will find the meeting very fruitful and use this event as a platform to meet new friends.

I would like to thanks all technical committee members for their contributions and supports for the success of ICCEE 2019.

I wish to announce that the organizer has planned for the next edition of ICCEE which will be taken place in Melaka in April 2020.

Have a wonderful day!

Prof. Zahriladha Zakaria, Ph.D.
Faculty of Electronic and Computer Engineering
Universiti Teknikal Malaysia Melaka (UTeM)
Melaka, Malaysia

Preface

This proceeding presents an abstract of research works in computer science, electrical and electronic engineering.

All manuscripts were presented during International Conference on Computer Science, Electrical and Electronic Engineering (ICCEE 2019) which was held at Seri Pacific Hotel Kuala Lumpur, Malaysia, on 29 and 30 April 2019. The editors of the proceeding would like to express the utmost gratitude and thanks to all reviewers in the technical committee for making this proceeding a success.

Melaka, Malaysia

Zahriladha Zakaria
Rabiah Ahmad

Contents

Lotus-G: The PVT TLM Virtual Platform for Early RUMPS401 Software Development

Arya Wicaksana

Abstract The fascinating growth of Multi-Processor System-on-Chip (MPSoC) has brought a daunting task to software programmers. The task is to build software which could completely utilize hardware potentials. The virtual platform is one of the solutions which allows software programmers to develop the software early before the hardware is available. As a demonstration of concept and viability, here the functional accurate programmer's view with time (PVT) transaction-level modeling (TLM) virtual platform: Lotus-G is presented here and used for early software development of a real MPSoC project: RUMPS401. The Lotus-G mimics the RUMPS401 that is an ultra-low power MPSoC project developed in Universiti Tunku Abdul Rahman (UTAR) VLSI Design Center, Malaysia. The virtual platform is developed using SystemC and transaction-level modeling (TLM) methodology with the use of additional high-level models from Open Virtual Platforms (OVP). The simulation of Lotus-G takes place in SystemC environment with the help of an instruction set simulator from OVP to execute the software on the platform. This research showcases the results and details of early MPSoC software development using a PVT TLM Virtual Platform with OVP simulator and high-level models for real MPSoC project.

Keywords Early software development · MPSoC · PVT · TLM · Virtual platform

1 Introduction

Software development of Multi-Processor System-on-Chip (MPSoC) has been a daunting task for software programmers as explained in [1]. The inclusion of many processors, IPs, peripherals, and interconnect architecture such as Network-on-Chip (NoC) makes an MPSoC today so complex. It is often added with the long existing challenges such as skyrocketing fabrication cost and tight time-to-market window

A. Wicaksana (✉)
Universitas Multimedia Nusantara, Tangerang 15810, Indonesia
e-mail: arya.wicaksana@umn.ac.id

Z. Zakaria and R. Ahmad (eds.), *Advances in Electronics Engineering*,
Lecture Notes in Electrical Engineering 619,
https://doi.org/10.1007/978-981-15-1289-6_1

as in [2]. Therefore it is essential for the software and hardware to be able to work together and to function as a single system for the first time (first-time success). This is why early software development also plays a crucial role as mentioned in [3]. The Field Programmable Gate Array (FPGA) and hardware prototype have been used for decades to do the early software development as described in [4]. However, the complexity of design and verification of an MPSoC today makes those solutions no longer viable to fit in. It is then imperative that a new approach is needed such as the Electronic System Level (ESL) design methodology as presented in [5].

The ESL design methodology proposes a virtual prototyping platform as a solution for enabling early software development as in [6]. The key concept is to raise the design abstraction level to the level where implementation details are not required hence they can be put aside. Details such as timing delay and power consumption are not the main concerns in early software development. In early software development, software programmers expect a representation of the targeted system to be available for them in order to run a simulation and debug the software. Here the virtual platform has to mimic the functionalities of the intended hardware system precisely. Thus, it is the Programmer's View abstraction level as shown in Fig. 1 suite the requirements for early software development. The other abstraction levels are suitable for other use-cases such as algorithm development, architectural exploration, and hardware verification as expressed in [7].

Lotus-G, as defined in [8], is a scalable MPSoC virtual prototyping platform built using SystemC and transaction-level modeling (TLM) methodology. As the superset of the C++ programming language, SystemC benefits software programmers to enter the MPSoC design league more swiftly compare to SystemVerilog and other hardware description language (HDL) or hardware verification language (HVL). The vast availability of C programmers in the world is also an advantage for the adoption of SystemC. The use of instruction set simulator (ISS) here is

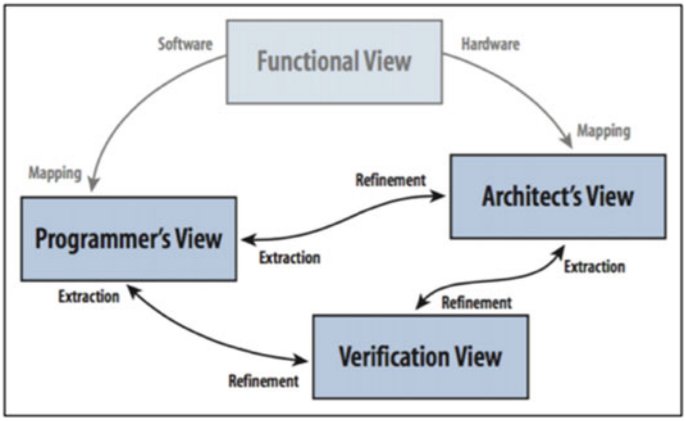


Fig. 1 TLM virtual platform abstraction levels scheme as in [7]

necessary to enable simulation of an executable binary file (software) on the Lotus-G virtual platform. The simulation is under SystemC environment and the ISS from Open Virtual Platforms (OVP) also conforms with the TLM loosely-timed (LT) coding style. Since the Lotus-G mimics the RUMPS401 MPSoC, thus the developed software could be run successfully on the system. This opens up possibilities for more integration and exploration using the virtual platform in terms of system architecture, performance, and optimization which are not limited to RUMPS401 only.

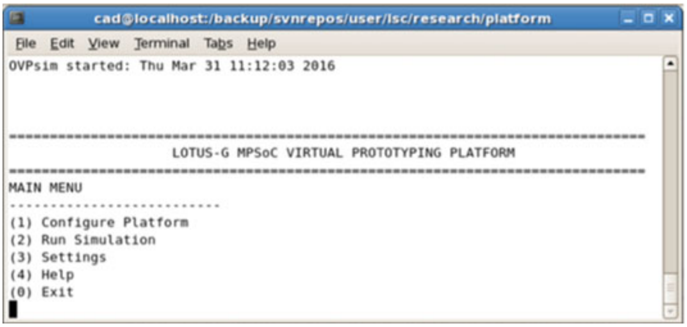
2 Material and Method

The material and method used to develop the Lotus-G virtual platform are the SystemC 2.3.1 with TLM 2.0.1, under the IEEE standard 1666-2011. OVP high-level models are also used in the virtual platform such as the processor, Bus, and generic memory models (ROM and RAM). The implementation details of the virtual platform follow the programmer's view with time (PVT) abstraction level with TLM loosely-timed (LT) coding style. Another key material used in this research is the RUMPS401 specification document that serves as the golden reference model for the system specification of the Lotus-G. Instruction set simulator (ISS) from OVP is also used to simulate the developed software on the virtual platform in SystemC with TLM environment.

2.1 Lotus-G

The Lotus-G is a scalable and functionally accurate programmer's view virtual platform with time (PVT). The programmer's view abstraction level satisfies the need of software programmers in developing and testing their software early. The abstraction level could be achieved by using TLM loosely-timed coding style as in [9], which also allows high-level TLM models to annotate timing delay and pass it along with the transaction. This timing delay is an optional feature as it can be ignored and will neither disrupt the simulation nor the behavior of the virtual platform. Nevertheless, the timing feature is still provided in Lotus-G and can be used optionally during the simulation, this allows timing estimation and optimization of the software.

The ARM Cortex-M0 processor model is used in Lotus-G and it is from the OVP (Open Virtual Platforms) [10]. The processor model is provided in a high-level language and it is able to connect to a SystemC–TLM virtual platform. High-level and fast simulator, the OVPFastSim is also used to simulate the software with the chosen processor model. Generic high-level memory models such as RAM and ROM (Intel Flash) are also used from the OVP. The interconnection between all of the processors and its attending subsystem is managed using NoC


cad@localhost:/backup/svnrepos/user/lsc/research/platform
File Edit View Terminal Tabs Help

```
OVPsim started: Thu Mar 31 11:12:03 2016



==========================================================================
                   LOTUS-G MPSoC VIRTUAL PROTOTYPING PLATFORM
==========================================================================
MAIN MENU
-------------------------
(1) Configure Platform
(2) Run Simulation
(3) Settings
(4) Help
(0) Exit
```

Fig. 2 Lotus-G virtual platform main menu display as in [12]

(Network-on-Chip) with 2D Mesh topology. Other IPs are also included in the Lotus-G in order to fully mimic the RUMPS401 features and behaviors.

The software debugging process is an important part in software engineering and it is supported by the Lotus-G via the GDB (GNU Project Debugger) [11]. Here software programmers are allowed to use GDB to debug their RUMPS401 software by connecting the GDB to the port that the Lotus-G provided during the simulation. After that point, the software programmers have full control over the execution flow of the executed program which is very helpful for debugging software. Figure 2 shows the CLI of the Lotus-G virtual platform. Menu (1) is required for software programmers to set up the system architecture of the virtual RUMPS401. As the virtual platform is scalable in terms of the number of processing elements (PEs) available. The connection between those PEs is managed by high-level models of NoC routers in a 2D-Mesh architecture. Menu (2) simulates the virtual platform with the software and hardware that have been set up in Menu (1) and the simulation runs with the help of a system level test bench. Menu (3) provides the ability for the users to change the simulation environment, for instance, the global quantum size.

2.2 *RUMPS401*

The RUMPS401 is an ultra-low power MPSoC packs up four processing elements (PEs) with each of the PEs packs an ARM Cortex-M0 processor, SST embedded flash, ARM’s RAM, AES Accelerator, and a DMA Controller. The project is initiated by Universiti Tunku Abdul Rahman VLSI Design Center in Malaysia. The fabrication is made possible in collaboration with Silterra under multi-project wafer (MPW) scheme. The RUMPS401 specification also sets one of the PEs for doing digital signal processing, therefore a MAC Accelerator is inserted in that particular PE. Figure 3 displays the RUMPS401 MPSoC and also the physical layout of the

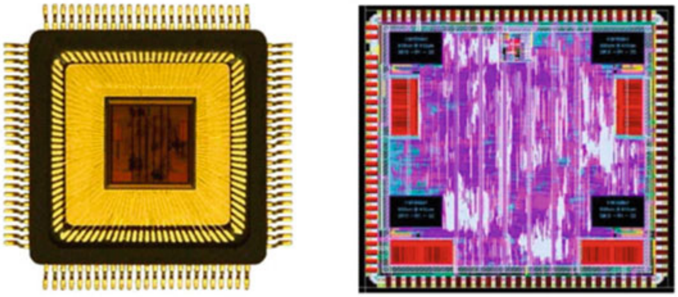


Fig. 3 UTAR RUMPS401 MPSoC (left) and the physical layout (right) as in [13]

RUMPS401. The system specification and architecture of RUMPS401 is mimicked by the Lotus-G virtual platform to provide an adequate level of abstraction for early software development.

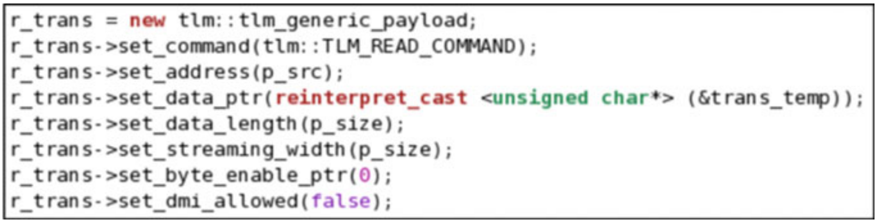
The RUMPS401 applications are not limited to AES encryption only. As there are plenty of applications that can utilize the RUMPS401 such as software-defined radio, smart meter, robotics, Internet-of-Things, and low power embedded systems. AES-128 encryption is chosen as the application here for demonstrating the early RUMPS401 software development using the Lotus-G PVT TLM virtual platform.

2.3 SystemC and TLM

The SystemC language with TLM provides ways for creating and managing process and communication. The language supports simulation of the prototyped hardware system. The chosen TLM coding style for the development of the Lotus-G virtual platform is loosely-timed rather than the approximately-timed, this is to allow software simulation to run faster.

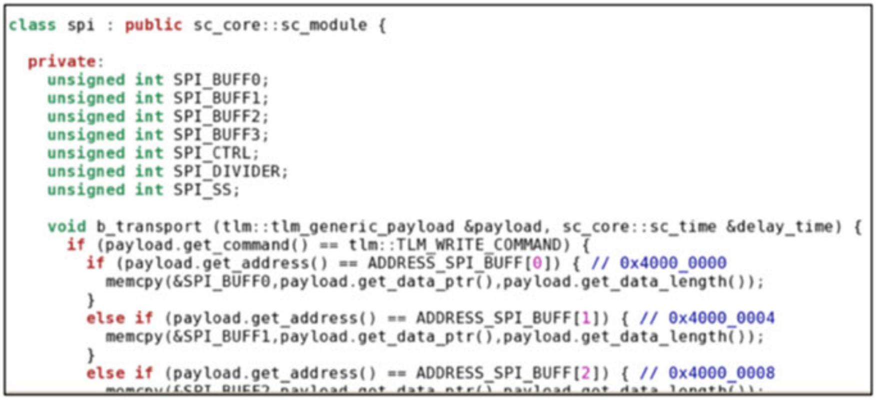
In TLM, the process could be categorized into three: initiator, interconnect, and target. The categorization is based on the module behavior towards the transaction. The initiator generates a transaction object and optionally put initial timing delay to the transaction. The interconnect receives transaction either from the initiator or the target and pass the transaction along without doing any modification to it except for giving additional timing delay. The target receives a transaction from the interconnect and processes it accordingly. Figure 4 shows the creation of a transaction object including the setting of its parameter.

The communication between each process in TLM is supported by passing along transactions using socket. Each of the SystemC modules that use TLM to communicate must implement socket to send and receive transactions. The target socket must be bound to a blocking transport callback function to provide the intended functionality upon receiving a transaction. This transaction also carries timing



```
r_trans = new tlm::tlm_generic_payload;
r_trans->set_command(tlm::TLM_READ_COMMAND);
r_trans->set_address(p_src);
r_trans->set_data_ptr(reinterpret_cast <unsigned char*> (&trans_temp));
r_trans->set_data_length(p_size);
r_trans->set_streaming_width(p_size);
r_trans->set_byte_enable_ptr(0);
r_trans->set_dmi_allowed(false);
```

Fig. 4 Creating transaction object



```
class spi : public sc_core::sc_module {

  private:
    unsigned int SPI_BUFF0;
    unsigned int SPI_BUFF1;
    unsigned int SPI_BUFF2;
    unsigned int SPI_BUFF3;
    unsigned int SPI_CTRL;
    unsigned int SPI_DIVIDER;
    unsigned int SPI_SS;

    void b_transport (tlm::tlm_generic_payload &payload, sc_core::sc_time &delay_time) {
      if (payload.get_command() == tlm::TLM_WRITE_COMMAND) {
        if (payload.get_address() == ADDRESS_SPI_BUFF[0]) { // 0x4000_0000
          memcpy(&SPI_BUFF0,payload.get_data_ptr(),payload.get_data_length());
        }
        else if (payload.get_address() == ADDRESS_SPI_BUFF[1]) { // 0x4000_0004
          memcpy(&SPI_BUFF1,payload.get_data_ptr(),payload.get_data_length());
        }
        else if (payload.get_address() == ADDRESS_SPI_BUFF[2]) { // 0x4000_0008
          memcpy(&SPI_BUFF2,payload.get_data_ptr(),payload.get_data_length());
```

Fig. 5 Implementation of callback function

information. When a process wants to communicate with another process, it calls the b_transport function of the target's socket. The call to the function is non-blocking and additionally, the caller may also call wait function to add the timing delay received to the global simulation time. Figure 5 shows the b_transport function as the implementation of the blocking transport callback function in one of the Lotus-G peripheral modules.

In Lotus-G, the initiator modules are the processor and the DMA Controller. The processor model is ARM Cortex-M0 from OVP and the DMA Controller is developed in high-level. A transaction object could be created as shown in Fig. 4. The interconnect modules in the virtual platform are the Bus and NoC. The Bus model is used from OVP and the NoC is developed in high-level following the functional specification of the RUMPS401 NoC. The rest of the hardware peripherals in the virtual platform are created as the target.

The programmer's view abstraction level strict implementation detail of the models to only functionally accurate. The timing is then modeled loosely in the virtual platform. There is no need to implement clock and to connect pins as in register transfer level (RTL). The main concern is the exact functionality and behavior of the hardware models to be mimicked precisely in high-level (PVT TLM).

This is very important to allow software programmers to write and test their software early by using the virtual platform.

One novelty that the Lotus-G virtual platform exhibit in providing timing approximation is the clock frequency feature. In the simulation menu, the user could set up the clock frequency to value and the value will be calculated and converted into a timing delay for the simulation. Thus, software programmers might be able to estimate roughly the time consumed by the software as the clock speed of the targeted hardware system is taken into account. Although this might limit simulation speed in overall, the timing estimation might still be useful to perceived by software programmers in such early stage of design and implementation. Further optimization could also be done in the software to speed up simulation time.

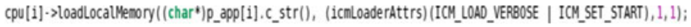
2.4 Instruction Set Simulator

An instruction set simulator (ISS) is required to simulate the executable binary file of the developed software. The developed software built using C programming language has to be cross-compiled to the same processor instruction set architecture. The Lotus-G virtual platform uses OVPFastSim as the ISS and it is provided by OVP.

The ISA of the Cortex-M0 processor is ARMv6-M Thumb as described in [14]. This information is required for the software to be cross-compiled correctly to match with the ISA of the targeted processor model. In this research, the software is cross-compiled using Mentor Graphic Sourcery CodeBench Lite Edition for bare metal as in [15]. The bare metal option is chosen as the RUMPS401 does not use an operating system to operate.

2.5 Software Bootloader

In RUMPS401 hardware design, a software bootloader is required to write program code to the internal program memory. The program code is obtained by the Bootloader circuit from an external source. The bootloader circuit is also designed to hold the activity of the processor core during the program loading process. In a high-level simulation environment this bootloader is not necessary. Instead, the feature is supported by loadLocalMemory function call provided by OVP as shown in Fig. 6.



```
cpu[i]->loadLocalMemory((char*)p_app[i].c_str(), (icmLoaderAttrs)(ICM_LOAD_VERBOSE | ICM_SET_START),1,1);
```

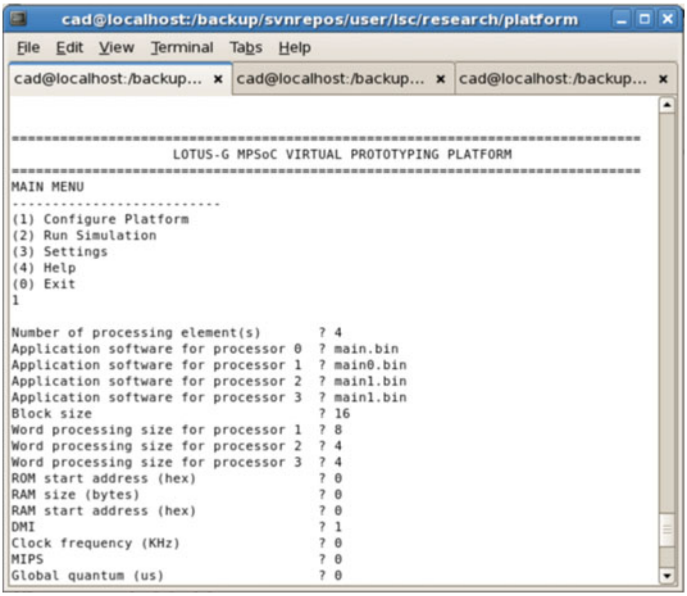
Fig. 6 Loading software to a processor

3 Results and Discussion

In early software development, the system partitioning process must already finished and produced system functionalities. System functionalities are then derived by the system architect to hardware and software. Here the encryption process is chosen to be done by the hardware while the management of the parallel processing and synchronization work are done by the software. RUMPS401 software in a whole is consisted of three files: application program file, interrupt service routine file, and linker file.

After getting all of the materials ready, all of the software files have to be cross-compiled for ARM Cortex-M0. After the cross-compilation is done, the executable file is produced and can be stored as a binary file. The executable binary file is then loaded to the Lotus-G and assigned to the respective processor. Since the Lotus-G is scalable, the processor is named using a number starting from 0 to n, the process is shown in Fig. 7.

The DMI stands for direct memory interface and when activated will allow the processor to directly access certain memory region without having to generate any TLM transaction. This feature is useful for fastening the simulation speed. The clock frequency is a unique feature in Lotus-G virtual platform where software programmers could set the clock frequency of the targeted system. The number value is in kHz and when given the Lotus-G will add timing delay into the transaction during the simulation. This unique feature gives software programmers the ability to estimate the time consumed by the software, hence further optimization can be done on the software. The MIPS stands for million instructions per second

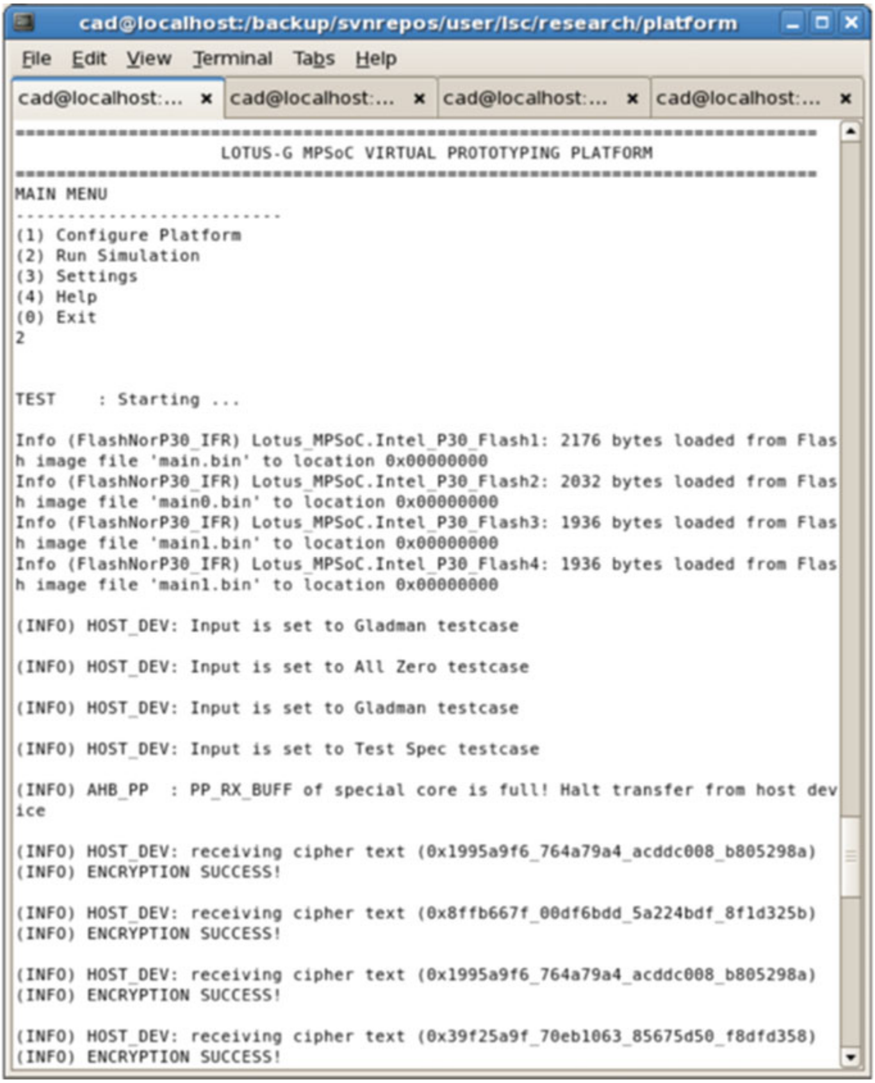

cad@localhost:/backup/svnrepos/user/lsc/research/platform
File Edit View Terminal Tabs Help
cad@localhost:/backup...
cad@localhost:/backup...
cad@localhost:/backup...

```
================================================================================
                      LOTUS-G MPSoC VIRTUAL PROTOTYPING PLATFORM
================================================================================
MAIN MENU
--------------------------
(1) Configure Platform
(2) Run Simulation
(3) Settings
(4) Help
(0) Exit
1

Number of processing element(s)       ? 4
Application software for processor 0  ? main.bin
Application software for processor 1  ? main0.bin
Application software for processor 2  ? main1.bin
Application software for processor 3  ? main1.bin
Block size                            ? 16
Word processing size for processor 1  ? 8
Word processing size for processor 2  ? 4
Word processing size for processor 3  ? 4
ROM start address (hex)               ? 0
RAM size (bytes)                      ? 0
RAM start address (hex)               ? 0
DMI                                   ? 1
Clock frequency (KHz)                 ? 0
MIPS                                  ? 0
Global quantum (us)                   ? 0
```

Fig. 7 Configuring the platform for a simulation

and is solely used for the OVPFastSim purpose. It decides how many million instructions have to be executed during the simulation within a second. The global quantum is a feature in TLM which manages the parallelization time for all threads during the simulation. Every single TLM initiator module is realized as a thread, thus the parallelization must be managed in terms of the length of time given by the simulator to each of the threads during simulation. This also includes the synchronization point of all threads. Figure 8 shows the simulation process after the setup has been completed.


cad@localhost:/backup/svnrepos/user/lsc/research/platform
File Edit View Terminal Tabs Help
cad@localhost:...
cad@localhost:...
cad@localhost:...
cad@localhost:...

```
==============================================================================
                       LOTUS-G MPSoC VIRTUAL PROTOTYPING PLATFORM
==============================================================================
MAIN MENU
--------------------------
(1) Configure Platform
(2) Run Simulation
(3) Settings
(4) Help
(0) Exit
2


TEST     : Starting ...

Info (FlashNorP30_IFR) Lotus_MPSoC.Intel_P30_Flash1: 2176 bytes loaded from Flas
h image file 'main.bin' to location 0x00000000
Info (FlashNorP30_IFR) Lotus_MPSoC.Intel_P30_Flash2: 2032 bytes loaded from Flas
h image file 'main0.bin' to location 0x00000000
Info (FlashNorP30_IFR) Lotus_MPSoC.Intel_P30_Flash3: 1936 bytes loaded from Flas
h image file 'main1.bin' to location 0x00000000
Info (FlashNorP30_IFR) Lotus_MPSoC.Intel_P30_Flash4: 1936 bytes loaded from Flas
h image file 'main1.bin' to location 0x00000000

(INFO) HOST_DEV: Input is set to Gladman testcase

(INFO) HOST_DEV: Input is set to All Zero testcase

(INFO) HOST_DEV: Input is set to Gladman testcase

(INFO) HOST_DEV: Input is set to Test Spec testcase

(INFO) AHB_PP   : PP_RX_BUFF of special core is full! Halt transfer from host dev
ice

(INFO) HOST_DEV: receiving cipher text (0x1995a9f6_764a79a4_acddc008_b805298a)
(INFO) ENCRYPTION SUCCESS!

(INFO) HOST_DEV: receiving cipher text (0x8ffb667f_00df6bdd_5a224bdf_8f1d325b)
(INFO) ENCRYPTION SUCCESS!

(INFO) HOST_DEV: receiving cipher text (0x1995a9f6_764a79a4_acddc008_b805298a)
(INFO) ENCRYPTION SUCCESS!

(INFO) HOST_DEV: receiving cipher text (0x39f25a9f_70eb1063_85675d50_f8dfd358)
(INFO) ENCRYPTION SUCCESS!
```

Fig. 8 Simulation process of AES-128 encryption software on Lotus-G

The simulation shown above is run using input driven from the outside that is called the HOST_DEV (test bench). The software then takes the input (test case: Gladman, All Zero, and Test Spec) and process it all together with the hardware (the Lotus-G). After the encryption process is done, the output is sent back to the test bench to be verified. The early software development carried on using the Lotus-G can produce up to six software with different hardware processor configurations ranging from 4 to 12 processors and the processing block size of the encryption varying from 16 to 512 blocks. There are two types of application software developed here, the first one (App 1) is to manage the distribution of the blocks and the second one (App 2) is to manage the encryption part. The degree of parallelism used is coarse grain and the distribution for each of the blocks to all of the processors is summarized in Table 1.

Simulation time statistics are also displayed after the simulation has ended as shown above from one of many Lotus-G simulations. Each of the CPUs' type, nominal MIPS (million instructions per second), final program counter, simulated instructions, and simulated MIPS are also displayed. The simulated time is the simulation duration that is independent to load and network delays of the host computer. User time is the wall-clock time given to OVPFastSim and system time is the time for performing system task for OVPFastSim process. The elapsed time shall be equal to the wall-clock from the beginning until the end of the simulation.

Table 1 Summary of early RUMPS401 software development

Block size	4 processors		8 processors		12 processors	
	App 1	App 2	App 1	App 2	App 1	App 2
16	8	4 × 2	4 × 4	–	4 × 4	–
32	12 × 2	8	8	4 × 6	4 × 8	–
64	24	20 × 2	12 × 2	8 × 5	8 × 5	4 × 6
128	44 × 2	40	20 × 4	16 × 3	12 × 10	8
256	88	84 × 2	40	36 × 3	24 × 9	20 × 2
512	172 × 2	168	76 × 2	72 × 5	48 × 7	44 × 4

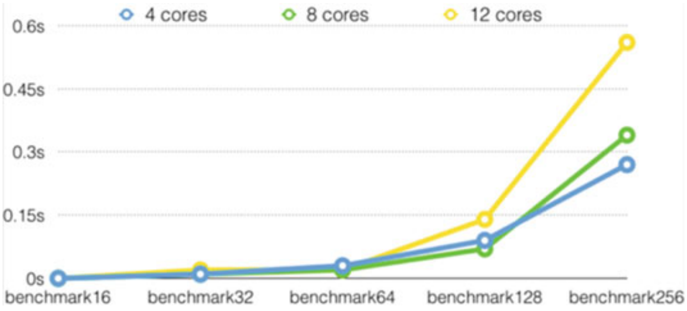


Fig. 9 Benchmark software simulation speed

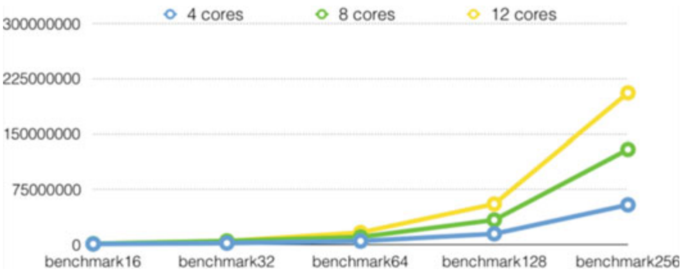
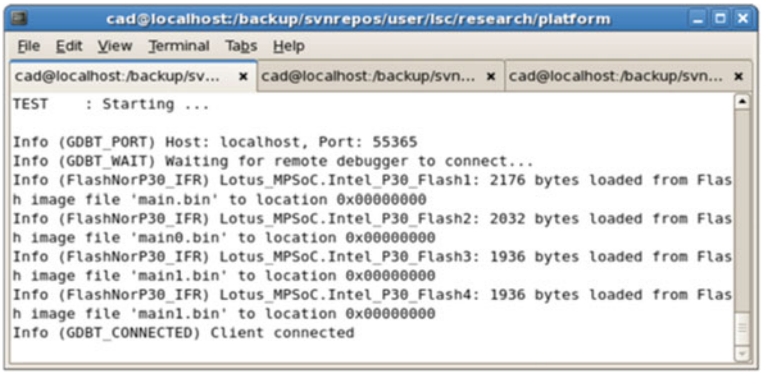


Fig. 10 Benchmark software simulated instructions


cad@localhost:/backup/svnrepos/user/lsc/research/platform
File Edit View Terminal Tabs Help
cad@localhost:/backup/sv... ×
cad@localhost:/backup/svn... ×
cad@localhost:/backup/svn... ×

```
TEST     : Starting ...

Info (GDBT_PORT) Host: localhost, Port: 55365
Info (GDBT_WAIT) Waiting for remote debugger to connect...
Info (FlashNorP30_IFR) Lotus_MPSoC.Intel_P30_Flash1: 2176 bytes loaded from Flas
h image file 'main.bin' to location 0x00000000
Info (FlashNorP30_IFR) Lotus_MPSoC.Intel_P30_Flash2: 2032 bytes loaded from Flas
h image file 'main0.bin' to location 0x00000000
Info (FlashNorP30_IFR) Lotus_MPSoC.Intel_P30_Flash3: 1936 bytes loaded from Flas
h image file 'main1.bin' to location 0x00000000
Info (FlashNorP30_IFR) Lotus_MPSoC.Intel_P30_Flash4: 1936 bytes loaded from Flas
h image file 'main1.bin' to location 0x00000000
Info (GDBT_CONNECTED) Client connected
```

Fig. 11 Debugging software simulation with GDB

The benchmark software is also developed specifically to test the Lotus-G virtual platform. All of the benchmark software runs successfully on the RUMPS401 MPSoC. The very same benchmark software also run successfully on the Lotus-G without having any modification. Figures 9 and 10 show the software simulation speed and simulated instructions of the benchmark software. The number after the benchmark is indicating the encryption block size. It can be figured out from the benchmark software simulation speed result comparison that more cores (processors) do not necessarily speed up the simulation time for all cases. Specifically for the benchmark128 where the 8 cores configuration is faster than the 4 and 12 cores configuration. These results are not only useful to software programmers in optimizing the software but also to system architect for doing architectural exploration.

The early developed software as shown in Table 1 including the benchmark software could be debugged during simulation in the virtual platform. As stated in Sect. 2, the Lotus-G here provides GDB connection for debugging the software during the simulation. Figure 11 shows the virtual platform is waiting for remote debugger to connect at port 55365. Software programmers could then use GDB client on other terminal and connect to the designated port. After the client has connected to the server, the simulation resumes and telling that client has been connected. On the GDB client terminal, software programmers have full control on the simulation and ability to debug the software per each instruction.

4 Conclusions

Early MPSoC software development could be achieved by using SystemC with TLM and an instruction set simulator (ISS). The ISS is used to simulate the instruction of the software. Thus, the benchmark software as shown here that are developed using C programming language are cross-compiled into the correct instruction set architecture of the processor. The memory mapping of the software is set in the linker file, this must be aligned with the hardware architecture. The hardware communicates back to the software via interrupt and this is handled by the interrupt service routine. The software also could control particular hardware peripheral using the driver.

All of the software files must be cross-compiled at once into a single executable binary file. Bootloader is also not required here to load the application software to the virtual platform for a simulation. It is also could be concluded that for early software development using a virtual platform, generic high-level models could be utilized without having to be exactly the same with the targeted hardware peripherals as long as the functionalities are still the same. The software simulation is faster than RTL simulation, and additional timing estimation is also provided. As described in this paper, the software programmers not only can develop and simulate their software but also debug the software using GDB.

All of the components presented here to support early software development is open-source and standard. The same software (benchmark software) used in RTL simulation of the RUMPS401 is runnable in the Lotus-G and vice versa. Further exploration such as the degree of task parallelization in the software could also be done including refinement of the virtual platform to approximately-timed TLM coding style or even cycle-accurate abstraction level for other use-cases than early software development.

Acknowledgements I would like to express my gratitude to Mr. Dicky Hartono and Mr. Chong-Ming Tang for their support and guidance on this research project. As well as Universiti Tunku Abdul Rahman (UTAR) for awarding me the research grant on RUMPS401 MPSoC MPW project and particularly this research work.

References

1. Ceng J, Sheng W, Castrillon J, Stulova A, Leupers R, Ascheid G et al (2009) A high-level virtual platform for early MPSoC software development. In: Proceedings of the 7th IEEE/ACM international conference on hardware/software codesign and system synthesis—CODES + ISSS '09 [Internet], pp 11–20. Available from: https://dl.acm.org/citation.cfm?doid=1629435.1629438 Accessed 8 May 2019
2. Jerraya A, Wolf W (2014) Multiprocessor systems-on-chips. Elsevier Science, Saint Louis
3. Martin G (2006) Overview of the MPSoC design challenge. In: Proceedings of the 43rd annual conference on design automation—DAC '06
4. Karyono, Wicaksana A (2013) Teaching microprocessor and microcontroller fundamental using FPGA. In: 2013 conference on new media studies (CoNMedia)
5. Rigo S, Azevedo R (2011) Electronic system level design. Springer, Dordrecht
6. Bailey B, Martin G, Piziali A (2007) ESL design and verification. Elsevier, Burlington
7. Kogel T, Braun M (2006) Virtual prototyping of embedded platforms for wireless and multimedia. In: Proceedings of the conference on design, automation and test in Europe, DATE [Internet]. Available from: https://www.researchgate.net/publication/221340948_Virtual_Prototyping_of_Embedded_Platforms_for_Wireless_and_Multimedia. Accessed 9 May 2019
8. Wicaksana A, Tang C (2017) Virtual prototyping platform for multiprocessor system-on-chip hardware/software co-design and co-verification. Comput Inf Sci 719:93–108
9. 2.3.4. Loosely timed approximately timed and untimed TLM [Internet] (2019). Available from: https://www.embecosm.com/appnotes/ean1/html/ch02s03s04.html. Accessed 9 May 2019
10. Welcome Page [Internet] (2019) Available from: http://www.ovpworld.org. Accessed 9 May 2019
11. http://www.ovpworld.org/documents/OVPsim_Debugging_Applications_with_GDB_User_Guide.pdf. Accessed 9 May 2019
12. Wicaksana A (2019) Multi-processor system-on-chip hardware and software modeling, co-design and co-verification [M.Eng.Sc.]. Universiti Tunku Abdul Rahman
13. Wicaksana A, Kusuma Halim D, Hartono D, Lokananta F, Lee S, Ng M et al (2019) Case study: first-time success ASIC design methodology applied to a multi-processor system-on-chip. Application specific integrated circuits—technologies, digital systems and design methodologies
14. ARM Information Center on ARM [Internet] (2019). Available from: http://infocenter.arm.com. Accessed 9 May 2019
15. Sourcery CodeBench [Internet] (2019). Available from: https://www.mentor.com/embedded-software/sourcery-tools/sourcery-codebench/overview. Accessed 9 May 2019

Utilization of Learning Management System (LMS) Among Instructors and Students

Salah Al-Sharhan, Ahmed Al-Hunaiyyan,
Rana Alhajri and Nabeil Al-Huwail

Abstract Today large number of universities around the world are equipped with Leaning Management Systems (LMS) to help in providing space for rich online learning environment, and to utilize its tools and functionalities to improve pedagogy and to increase the quality of learning. This paper aims to identify issues related to the utilization of LMS. Understanding these issues, allows developing better policies and systems to assist contributing to better learning experiences and academic success. The paper introduces a case study conducted to identify the utilization of LMS in the Gulf University for Science and Technology (GUST), the first private university in Kuwait. The study was conducted to investigate the degree of LMS utilization of functions and features, among instructors and students at GUST.

Keywords LMS · E-learning · ICT · Technology utilization

1 Introduction

Learning Management System (LMS) is an online portal that provides space for classroom resources, tools, and activities to be shared easily among instructors and students. LMS has a variety of applications and tools that motivate universities around the world to encourage faculty members to utilize them for teaching and learning practices [1], and to assist them to track students' activities in more manageable manner, allowing collaboration, involvement, and interaction [2]. LMS provides variety of functions and communication tools that agreed to support

S. Al-Sharhan
Computer Science Department, Gulf University for Science and Technology, Mubarak al-Abdullah, Kuwait

A. Al-Hunaiyyan (✉) · R. Alhajri · N. Al-Huwail
Computer Science Department, Public Authority of Applied Education and Training, Kuwait City, Kuwait
e-mail: hunaiyyan@hotmail.com

Z. Zakaria and R. Ahmad (eds.), *Advances in Electronics Engineering*,
Lecture Notes in Electrical Engineering 619,
https://doi.org/10.1007/978-981-15-1289-6_2

teaching and learning such as assignments, announcement, quizzes, discussion forum, chat, resources, and others [3]. These communication tools, which can be used either synchronous or asynchronous, not only enrich the teaching and the learning process, but also facilitate communication and collaboration among students and instructor [4]. It is emphasised by Makrakis and Kostoulas-Makrakis [5] that with LMS, learning can be shifted from instructors-centred learning to students-centred learning, and that instructors' role should focus on facilitator rather than just knowledge transmitter.

Riddell [6] deliberated twelve biggest names among LMS providers, such as Blackboard, Moodle, Desire2Learn, Sakai, Jenzabar, Pearson Learning Studio/ eCollege, Canvas by Instructure, Angel Learning, and Cengage Leaning/MindTap, Adrenna, McGraw-Hill Connect. According to [7], Blackboard is the most popular LMS in HE institutions in the USA. Blackboard represents 33% of popularity, second is Moodle (19%), then Canvas (17%). Moodle continues to be the second most popular LMS by a large number of institutions and remains very popular among smaller colleges in the USA and Canada. The present study highlights utilization issues of LMS in general, and introduced a case study for the Gulf University for Science and Technology (GUST). The case study aims to demonstrate the degree of LMS utilization among instructors and students at GUST. Understanding these issues, allows developing better policies and systems to assist contributing to better learning experiences and academic success.

The rest of this paper is organized as follows: Sect. 2 focuses on the utilization of LMS. Section 3 introduces the case study with the findings, while Sect. 4 concludes the study and suggests future directions.

2 LMS Utilization

LMSs have been carried out since 2003 in higher educational institutions for effective communication. Nowadays, LMSs are widely used in universities, and it is believed that they play an important role to achieve the pedagogical elements, if it is used correctly. However, little attention is paid to how well these systems are utilized in higher education [8]. With LMS usage, Azlim et al. [9] stressed that students and instructors have flexibility on collaboration and discussion through LMS functions and tools. However, they insisted that instructors need to be given the support to encourage students to be actively involved in the LMS. As reported by Dahlstrom et al. [10], instructors and students believe that LMS enhances the teaching and learning experiences. However, instructors and students rarely use the more advanced functions of LMS.

An interesting research done by Azlim et al. [9] aimed to identify the utilization of LMS in HE institutions in Malaysia among instructors. They used a quantitative approach, with a questionnaire distributed to 93 instructors. They examined some LMS functions and tools, such as Groups, Chat, Discussions, Exercises, Announcements, and Documents. Although instructors have positive perceptions towards the value and benefits of LMS, results showed a low percentage of utilization of LMS from instructors. Another study was conducted by Alghamdi and Bayaga [1] to understand the relationship between 222 instructors representing six universities in Saudi Arabia. The study investigated instructors' perceptions and attitudes towards LMS tools and functionalities. The findings revealed that LMS tools and applications were not actively used for most of the teaching courses, and indicated that older instructors (over 40 years) tended to use LMS functions more than the younger counterparts. However, the study indicated some barriers such as fear of usage.

It is also documented by Dahlstrom et al. [10] that user satisfaction is at its highest level for basic LMS features and lowest for features designed to foster collaboration and engagement. Therefore, system environments should integrate collaboration and interactivity features into the user experience, helping instructors to easily make these features an integral part of their courses. Mobile devices have become ubiquitous, allowing students to access systems, such as the LMS, are becoming more common and increasingly important. Todays' digital learning environments requires anytime/anywhere access to course materials and 24/7 collaboration and engagement by mobile friendly devices. In addition, personalization of the LMS settings and interface will certainly add value to the student experience by promoting success strategies within and across courses.

3 Case Study: GUST

3.1 LMS Utilization at GUST: An Investigation

The Gulf University for Science and Technology (GUST) is the first private university in Kuwait, with over 3600 students and 179 faculty members. Since its establishment, GUST deployed Information and Communication Technologies (ICT) as part of the infrastructure and established an e-learning center of excellence in 2005 [11]. They aimed at transforming learning and instructional forms in ways that extend beyond the efficient delivery. The E-learning Center of Excellence (ECE) was equipped with LMS to enrich the teaching and learning practices by the proper utilization of LMS management and communication tools. The investigation presented in the next section explores the utilization of Moodle, a learning management system, used at GUST. The findings are based upon the actual use of LMS by instructors and students for the academic year 2017–2018. Data were pulled from the LMS transaction records, and were carried out using log data obtained from various courses dispensed in GUST, indicating activities performed.

3.2 Findings

The results presented in this section include the actual LMS visits by students and instructors (Web-LMS and Mobile-LMS) as shown in Table 1. In addition, LMS functions created by instructors to facilitate management and communication activities among instructors and students are presented in Table 2 for the fall semester and in Table 3 for spring semester 2017–2018.

Table 1 shows that the utilization of LMS in spring semester is more than in the fall. Moreover, activities and visits to LMS by students and instructors from the College of Art and Science is more than the activities and visits in the College of Business during the academic year 2017/2018. This is related to the nature of the courses of the college of Art and Science, and to the high number of students and instructors in the college. Notably, LMS visits by web is more than by mobile devices. Mobile is rarely used to access LMS activities, which could be related to the limited feature of LMS through mobile devices. The growing learner needs, urge educational institutions to go mobile. When learning goes mobile, and the learning management systems become fully compatible with mobiles, LMS functions and tools should also be available on mobile devices [12]. This will help to provide learners anytime/anywhere access to the online courses, and to allow instructors to effectively manage the learning activities.

Table 2 shows the LMS functions and tools created by instructors for the fall semester 2017–2018. It is obvious that the activities by the college of Art and Science with (8392) are more than in the College of Business with (4242). In addition, the file function is more used by instructors, then assignments, and discussions. On the other hand, creating an interactive book, and chatrooms was not considered by instructors.

Similarly, Table 3, shows the LMS functions and tools created by instructors for the spring semester 2017–2018. As in the fall semester, the activities by the college of Art and Science is far more with (10,022), while only (4137) for the College of Business. The file function is more utilized by instructors, then comes assignment, URL, then discussions. On the other hand, creating an interactive book, and chatrooms also was not considered by instructors.

As shown in Tables 1, 2, and 3, and Fig. 1, the utilization of LMS functions and tools varies according to the purpose of use. Some functions are broadly used such as files and assignments, while other interactive tools such as creating a book and chatrooms are not used.

Table 1 Actual LMS visits (web LMS vs mobile LMS)

	LMS visits by web		LMS visits by mobile	
	Fall	Spring	Fall	Spring
College of Art and Science	436,235	509,866	6447	16,946
College of Business	296,873	311,127	4378	13,446
Total	733,108	820,993	10,825	30,392

Table 2 Number of LMS functions created by instructors—fall semester 2017–2018

	Assignments	Book	Chat	Folder	Forum-discussion	Hot pot quiz	Label	Quiz	File	Turn it in	URL	Total
College of Art and Science	1415	13	2	305	525	49	424	230	4327	436	666	8392
College of Business	438	0	0	154	280	0	139	275	2860	6	90	4242
Total	1853	13	2	459	805	49	563	505	7187	442	756	12,634

Table 3 Number of LMS functions created by instructors—spring semester 2017–2018

	Assignments	Book	Chat room	Folder	Forum-discussion	Hotpot quiz	Label	Quiz	File	Turn it in	URL	Total
College of Art and Science	1298	7	0	280	516	61	518	570	5182	596	993	10,022
College of Business	365	0	0	215	395	0	157	152	2820	0	33	4137
Total	1663	7	0	495	911	61	675	722	8002	596	1026	14,159

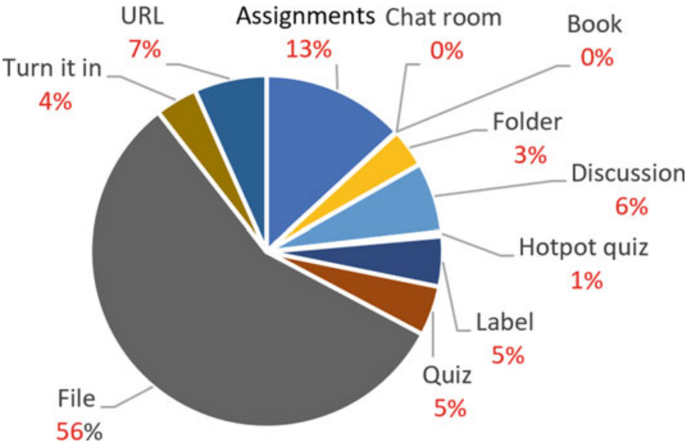


Fig. 1 LMS functions created by instructors—academic year 2017/2018

As illustrated in Fig. 1, it is noted that the file function created by instructors represents 56% of the total functions and tools provided by the LMS, while interactive features are less handled by instructors.

This finding is in line with [13], that today's LMS is only used to focus on the delivery of learning materials rather than learning the proper techniques. In addition, Almarashdeh et al. [14] stated that LMS includes many administrative, collaborative, assessment, and pedagogical elements that support and advance the learning process and help in distributing the course materials at a distance. Having this pointed out, Chung [8] stressed that LMS is not just a platform to only deliver learning resources, but it must be properly utilized so that it becomes a great venue for collaborative learning activities. Figure 1 shows the percentage of functions created by instructors during the academic year 2017/2018.

4 Conclusion and Future Directions

Learning Management System (LMS) is an essential tool for university students because it helps them to keep up with their coursework, get instant notifications regarding exams, quizzes, and their daily assignments. Similarly, instructors have an easier time reaching out to their students out of class hours and can instantly update them with coursework issues. Although LMS is widely used in universities and offer many advantages in education, Azmi et al. [15] believe that there is a debate about its effectiveness in education. LMS demands lots of responsibilities and requires technical skills, and virtual insights from instructors due to the technical nature of LMS [16].

This paper highlighted several issues related to the utilization of LMS functions and features, and reported some studies from various HE institutions. The paper then presented a case study that includes an investigation conducted at the Gulf University for Science and Technology (GUST) in Kuwait, to identify the utilization of LMS, and to understand the degree of LMS utilization of functions and tools among instructors and students. The results indicate a low percentage of the utilization of LMS functions, in which web-based LMS is more utilized than mobile-based LMS. Administrative and management functions of LMS is usually utilized, while interactive learning functions, such as chat and interactive book, are rarely handled by instructors. Understanding these issues, allows developing better policies and systems to assist contributing to better learning experiences and academic success. To accomplish that, it is essential to encourage the utilization of LMS, and to focus on learning activities to achieve the pedagogical elements, rather than just administration of the course. It is argued by Daniels et al. [17] that LMS itself is not the total solution to the engagement of students in teaching and learning, they stress that instructors play important role to encourage students to get benefits from LMS tools and functions. Thus, universities should provide proper training and guidance for students and lecturers using LMS functions and tools.

This initial study will be considered as a base for the next study that aims to focus on the perceptions and attitudes towards LMS usage, functions, and capabilities at GUST. More investigation is also required to highlight several issues related to demographic element that impacts LMS optimizations. This can lead us to construct a framework as guidance for instructors using tools in LMS to create active learning activities, better course administration, and effective collaborations among instructors and students. The future study can also spot the lights on the barriers that may affect LMS utilization at GUST.

Acknowledgements The authors of this paper would like to thank Kuwait Foundation for the Advancement of Sciences (KFAS), to their generous financial support for our participation in the *International Conference on Computer Science, Electrical and Electronic Engineering (ICCEE 2019), 29th to 30th April 2019, Kuala Lumpur, Malaysia*. Thanks go also to the persons working in the Research Directorate for facilitating this support.

References

1. Alghamdi S, Bayaga A (2016) Use and attitude towards Learning Management Systems (LMS) in Saudi Arabian universities. Eurasia J Math Sci Technol Educ 12(9):2309–2330
2. Emelyanova N, Voronina E (2014) Introducing a learning management system at a Russian university: students' and teachers' perceptions. Int Rev Res Open Distance Learn 15(1): 272–289
3. Bacow L, Bowen W, Guthrie K, Lack K, Long M (2012) Barriers to adoption of online learning systems in US higher education. Ithaka S+R, New York. Available from: http://www.sr.ithaka.org/publications/barriers-to-adoption-of-online-learning-systems-in-u-s-higher-education/. Accessed 10 Dec 2017

4. Venter P, Rensburg M, Davis A (2012) Drivers of learning management system use in a South African open and distance learning institution. Australas J Educ Technol 28(2):183–198
5. Makrakis V, Kostoulas-Makrakis N (2012) The challenges of ICTs to online climate change education for sustainable development: the ExConTra learning paradigm. In: Proceedings of the 5th conference on elearning excellence in the Middle East—sustainable innovation in education, Dubai, UAE, 30 Jan to 2 Feb 2012
6. Riddell R (2013) 12 learning management system providers and what they bring to classrooms. EducationDIVE, REPORT. Available from: https://www.educationdive.com/news/12-learning-management-system-providers-and-what-they-bring-to-classrooms/97613/. Accessed Dec 2017
7. Edutechnica (2016) 4th annual LMS data update. Edutechnica. Available from: http://edutechnica.com/2016/10/03/4th-annual-lms-data-update/. Accessed Dec 2017
8. Chung C (2013) Web-based learning management system considerations for higher education. Learn Perform Q 1(4):24–37
9. Azlim M, Husain K, Hussin B, Zulisman M (2014) Utilization of learning management system in higher education institution in enhancing teaching and learning process. J Hum Cap Dev 7(1)
10. Dahlstrom E, Brooks D, Bichsel J (2014) The current ecosystem of learning management systems in higher education: student, faculty, and IT perspectives. Research report, ECAR, Louisville, CO. Available from: http://www.educause.edu/ecar
11. Al-Doub E, Goodwin R, Al-Hunaiyyan A (2008) Students' attitudes toward e-learning in Kuwait's higher education institutions. In: Proceeding of the 16th international conference on computers in education (ICCE 2008), Taipei, Taiwan, 27–31 Oct 2008
12. Kumar A (2017) Make your LMS mobile compatible in 4 easy ways [online]. Available from: https://blog.commlabindia.com/elearning-design/4-ways-to-make-lms-mobile-compatible
13. Christie M, Jurado R (2009) Barriers to innovation in online pedagogy. Eur J Eng Educ 34 (3):273–279
14. Almarashdeh A, Sahari N, Zin N, Alsmadi M (2010) The success of Learning Management System among distance learners in Malaysian universities. J Theor Appl Inf Technol 80–91
15. Azmi M, Zeehan S, Fahad S, Maryam F, Hisham A (2012) Assessment of student perceptions towards e-learning management system (E-LMS) in a Malaysian pharmacy school: a descriptive student. Malays J Public Health Med
16. Kanninen E (2008) Learning style and e-learning. Tampere University of Technology
17. Daniels J, Jacobsen M, Varnhagen S, Friesen S (2013) Barriers to systematic, effective, and sustainable technology use in high school classroom. Can J Learn Technol 39(4)

Optical Amplification in Multiple Cores of Europium Aluminium Composite Incorporated Polymer-Based Optical Waveguide Amplifier by Using Mosquito Method

Nur Najahatul Huda Saris, Azura Hamzah, Sumiaty Ambran, Osamu Mikami, Takaaki Ishigure and Toshimi Fukui

Abstract In this paper, Europium Aluminum Benzyl Methacrylate (Eu–Al/BzMA) optical waveguide amplifier is fabricated with multiple cores of 50 μm diameters by a unique procedure known as the Mosquito method which, utilized a micro-dispenser desktop machine. The waveguides which used the optimum dispensing criteria to fabricate desired core diameter with high precision are then optically tested. The highest optical gain of 1.08 dB/cm with 0.77 dBm/cm insertion loss has been demonstrated in this device at 617 nm amplification wavelength. Suppression of concentration quenching, high compatibility with other materials, ease of fabrication, and low cost make such rare-earth-metal ion-incorporated polymer waveguide amplifiers suitable for providing gain in many integrated optical devices in the communication area.

Keywords Optical amplification · Rare-earth-metal polymer composite · Waveguide amplifier · Mosquito method

N. N. H. Saris (✉) · A. Hamzah · S. Ambran · O. Mikami
Department of Electronic Systems Engineering, Malaysia-Japan International Institute of Technology, Universiti Teknologi Malaysia, 54100 Kuala Lumpur, Malaysia
e-mail: nnhuda3@live.utm.my

T. Ishigure
Faculty of Science and Technology, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama 223-8522, Japan

T. Fukui
Intelligence & Energy Materials Research Laboratory, KRI, Inc. 134, Chudoji Minami-machi, Shimogyo-ku, Kyoto 600-8813, Japan

Z. Zakaria and R. Ahmad (eds.), *Advances in Electronics Engineering*,
Lecture Notes in Electrical Engineering 619,
https://doi.org/10.1007/978-981-15-1289-6_3

1 Introduction

The demand for optical amplifiers has grown widely since the introduction of Erbium Doped Fiber Amplifier (EDFA) and it is implemented in many applications, such as telecommunication [1] and biomedical industry. Therefore, the optical amplifier is very important to increase the system performance in terms of bandwidth, whether for the applications of longer or shorter signal transmission. This is because the performance of the system may reduce due to the attenuation factor. To overcome this problem, the implementation of the optical amplifier becomes necessary in the optical transmission system.

Along the glass fiber development, there is an increasing interest of research in many other optical components for Polymer Optical Fiber (POF). Even though the attenuation of silica glass fiber is thousand times lower than that of POF, which is approximate to 0.2 and 0.2 dB/m for glass fiber and POF respectively, but, it is competitive in comparison with the glass fibers for local area networks (LANs) as one of the short distance application which may need fiber length of 10–20 m [2]. In fact, it has the advantages of low production cost and processing flexibility over inorganic optical fiber [3, 4]. Besides that, the polymer optical devices with low loss window can be doped with rare-earth organic complexes, in which some of them can amplify the visible light [5–7].

For the case of short distance communication, such as LANs, the amplification of light may be necessary in order to compensate for the loss that occurs in the optical devices, such as splitters in POF transmission by inserting compact optical waveguide amplifier incorporated polymer in the POF link. However, high optical amplification is difficult to achieve at the short length of RE-doped fiber due to the migration of the photon energy between the adjacent of the rare earth ion at high concentration, which usually refers as concentration quenching. To address this problem, the rare earth-metal (RE-M) composite is used. The features of the RE-M have been reported elsewhere [8–12].

In this work, Europium (Eu) and Aluminium (Al), represent the rare earth and metal, respectively, have shown the performance in terms of optical gain and insertion loss. Eu which emits in the red has its advantage whereby it is free from photodegradation compared to the organic dyes [5, 13]. Meanwhile, the organic dyes have a critical issue on the photodegradation due to the continual exposure to the pump light, which can cause the emission drops progressively [14]. On top of that, by implementing the RE-M incorporated polymer composite, it is possible to suppress the multiphoton relaxation by a heavy metal ion. The gain amplification of Eu–Al/BzMA with different concentration for 10 wt% has been reported to have 7.1 dB/cm by utilizing Variable Stripe Length (VSL) method [9]. However, the deployment of the VSL is impractical and complicated for the current application [15, 16]. By using the coaxial pumping, the Eu–Al/BzMA has shown 3.24 dB/cm optical gain amplification through 13 wt% concentration [16]. Both findings have been recorded from the core diameter of 100 μm. Hence, this research would focus on fabricating and testing the 50 μm core diameter of the waveguide as it is the typical diameter of the multimode waveguide by utilizing the single stage single pass (forward pump) configuration at 13 wt% core concentration.

Despite the range of considerable fabrication processes for GI multimode polymer waveguide, it was decided that the best method and procedure for the investigation in this research was the Mosquito method. It is based on the fact that it is straightforward and relatively fast to fabricate waveguides using fabrication facilities, which is Micro-dispenser desktop machine at Ishigure Info-Optics laboratory. Since the waveguide structure used is embedded planar, hence, no complex fabrication technique is needed. On top of that, the fabrication technique using a micro-dispenser machine is very unique since it is photomask free, besides, no chemical etching process and large UV exposure apparatus are needed [17]. It has the capability to create the GI multimode circular core directly on-board. To the best of author's knowledge, there is no comprehensive work performed in studying the Mosquito method dispensing criteria for XCL01 as a cladding monomer and Eu–Al/BzMA as a core monomer, respectively. Since Eu ions emit the fluorescence light around 617-nm wavelength ($^5D_0 \rightarrow {}^7F_2$) under excitation at 532-nm wavelength light, it can be considered as interesting wavelength for signal amplification in the integrated optics applications, e.g., data transmission in optical interconnects, in-vehicle network, and visible light communication application.

2 Mosquito Method as Fabrication Technique

Generally, the polymer waveguide core cross-section shapes are designed to be square or rectangular, and the refractive index of the core region is uniform. However, the polymer waveguide, which is studied in this research was Eu–Al/BzMA and it is considered to be a Rare Earth-Metal (RE-M) incorporated polymer. It is mainly having an embedded planar structure in combination with graded index optical fiber circular core; upon realizing the superiority of the GI multimode fiber in high-speed transmissions. GI circular core is important to reduce the coupling efficiency of the fiber to waveguide connection link. Hence, to fabricate such a polymer waveguide with multiple channels of circular cores formed by parabolic refractive index profiles, the waveguides are all fabricated by using a micro dispenser machine known as the Mosquito method.

A typical 50 µm multimode core diameter of fabricated waveguide sample with a separation distance of 250 µm, which is the typical optical fiber pitch, is illustrated in Fig. 1. It is comprising a cladding and circular core. The facet of the waveguide array is plotted in Fig. 1a with the highlighted cross-section and overview of the core along the waveguide in Fig. 1b and c, respectively. In this research, six channels of parallel GI multimode circular core are fabricated including the first core as a dummy core. Multiple channels were prepared in a waveguide for efficiency and experimental convenience.

In order to fabricate the waveguide, at first, the cladding monomer XCL01, is coated on the glass substrate. After that, the core monomer is dispensed directly into

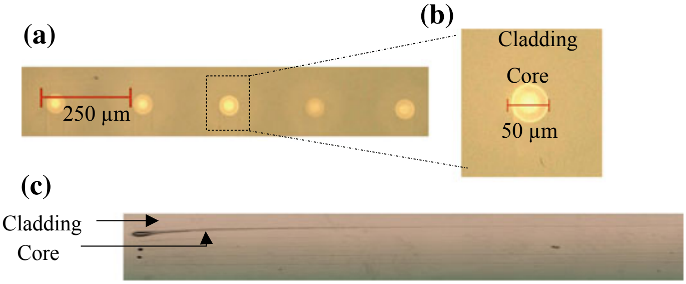


Fig. 1 **a** Sample of fabricated Eu–Al polymer optical waveguide on a glass substrate comprising of 50 µm waveguides with pitch of 250 µm; **b** cross-section of highlighted 50 µm circular core waveguide and **c** overview of the core along the waveguide

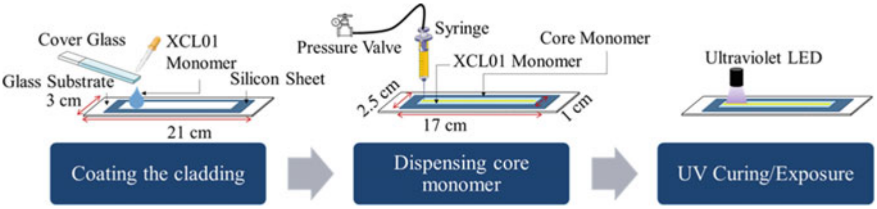


Fig. 2 Mosquito method fabrication steps

the viscous cladding monomer through the syringe that is attached to the micro-dispenser desktop machine. Then, the core is thermally polymerized soon after the cladding monomer is UV cured. The general view of the Mosquito method fabrication steps is illustrated in Fig. 2.

2.1 Optimum Dispensing Parameters

Previous studies by Takaaki Ishigure laboratory have reported that there was a tendency whereby the fabricated waveguide using Mosquito Method is affected by needle inner diameter, dispensing scanning speed, core or cladding monomer viscosity, and dispensing pressure, even though the material of cladding and core used were different from XCL01 and Eu–Al/BzMA incorporated polymer composite, respectively [2, 17, 18]. Hence, these three dispensing criteria were investigated to create the desired waveguide core characteristics using XCL01 and Eu–Al/BzMA as a cladding and core monomer, respectively.

For this purpose, five stainless steel needles (SUS304) provided by Musashi Engineering Inc. were used. Circular inner needle diameters of 110, 130, 150, 170, and 190 µm and dispensing scanning pressure ranging from 210 to 630 kPa were varied. The core and cladding viscosity and the concentration were fixed at 2000 cPs and 5000 cPs, respectively. On the other hand, the concentration of the core used also was remained constant at 13 wt%.

The dependency of the core diameter on these three parameters is shown in Fig. 3. Result plots in Fig. 3a show the core diameter versus dispensing scanning speed for variation of dispensing pressure with fixed inner and outer needle diameter of 150 μm and 300 μm, respectively. It is chosen since the previous work has focused on 150 μm inner needle diameter as a reference.

Next, the justification of the dispensing criteria, especially in inner needle diameter for Eu–Al/BzMA waveguide fabrication using Mosquito Method in this research was extended by varying the inner needle diameter ranging from 110 to 190 μm and the result is plotted in Fig. 3b. The plots show the average of five core diameters created on a waveguide. From the results, 325 kPa dispensing pressure with 150 μm inner needle diameter and 61.8 mm/s of dispensing speed is set to create 50 μm core diameter of the waveguide.

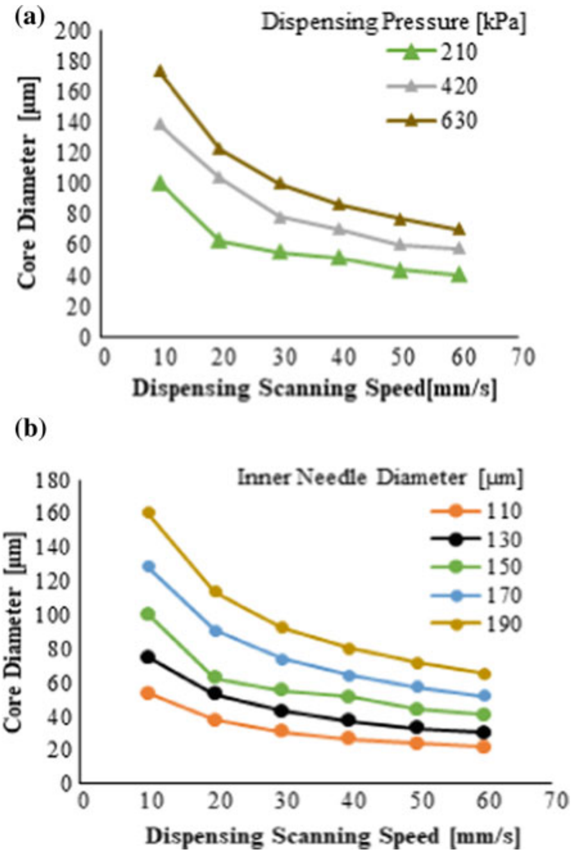


Fig. 3 Core diameter versus dispensing scanning speed for variation of **a** dispensing pressure with 150 μm inner needle diameter and **b** inner needle diameter with 210 kPa dispensing pressure

2.2 Circularity of the Core Diameter

The reproducibility of the core using the Mosquito method was then justified to confirm ±5% tolerance of core diameter as stated in [2]. Therefore, six waveguides with 50 µm core diameter were fabricated based on the appropriate dispensing conditions found for XCL01 and Eu–Al/BzMA, in which 150 µm inner needle diameter was used with 61.8 mm/s of drawing velocity and 325 kPa dispensing pressure. The cross-section of the fabricated core of each waveguide is shown in Table 1 and the average diameter is analyzed precisely by using the absolute approximation error, ϵ and relative error, n to calculate the percentage error. Then, the results are tabulated in Table 2. The used formulas are stated as follows:

Table 1 The core cross-section of each waveguide

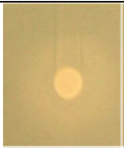
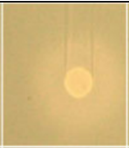
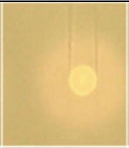
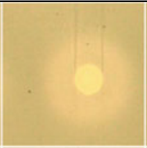
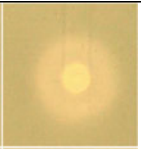
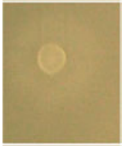
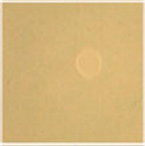
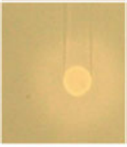
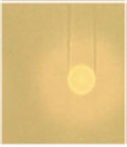
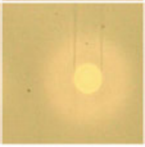
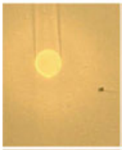
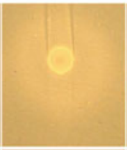
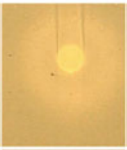
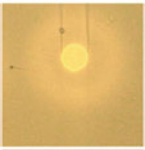
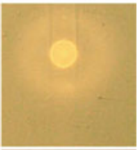
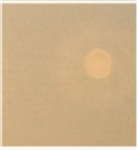
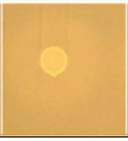
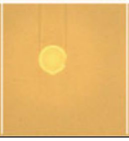
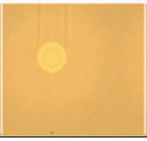
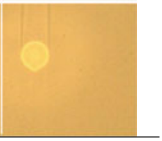
	Channel	1	2	3	4	5
Waveguide Number	1					
	2					
	3					
	4					
	5					
	6					

Table 2 The percentage error of average core diameters in each waveguide

Waveguide number	1	2	3	4	5	6
Average core diameters in each waveguide (μm)	53.67	49.61	51.65	50.19	49.57	51.63
Percentage error (%)	7.35	0.78	3.30	0.38	0.86	3.26

$$\epsilon = |Approximate\ value - Exact\ value| \tag{1}$$

$$n = \frac{\epsilon}{|Exact\ value|} \tag{2}$$

$$Percentage\ error = n \times 100\% \tag{3}$$

Based on the collected data, it is found that the tolerance of core diameters recorded by all waveguides was within the range from 47.5 to 52.5 μm. The values were corroborated the previous finding, except for core diameter of waveguide number 1. Contrary to expectation, the waveguide number 1 has shown 7.35% calculated percentage error. Since the previous experimental approximation is ±5% tolerance of core diameter, the waveguide number 1 was recorded 2.35% off. The reason for this rather contradictory result is still not entirely clear but it is believed to be due to the different interim time of diffusion for each core that can cause the fabricated core diameter to be unevenly created. The interim time refers to time delay or gap between the exposure standby time until the ultraviolet exposure is finished. Overall, the influence of this error on core diameter would appear to be due to "liquid state" core monomer, in which it is dispensed directly within the "liquid state" of cladding monomer. Therefore, the concentration distribution is dispersed unevenly. It should be noted that there might be some other influences such as humidity and temperature that affect the formation of the desired core diameter during the fabrication process that would result in the increased core diameter percentage error.

After examining the fabrication conditions for Eu–Al/BzMA, it is considered effective to change the four parameters to realize the desired structure in accordance with the basic Mosquito method using the dispenser described above.

3 Performance of Eu–Al/BzMA

The "back to back" experiment was conducted to measure the insertion loss in the fabricated waveguide, in which the waveguide core is launched through a 1 m of single-mode fiber (SMF) and received via a 2 m of 200-μm SI multimode fiber, respectively. The insertion loss data were recorded using a power meter. Then, the optical gain in the fabricated waveguide was measured by using forward pumping setup that replaced 1 m of SMF with the 50 μm GI multimode coupler at the

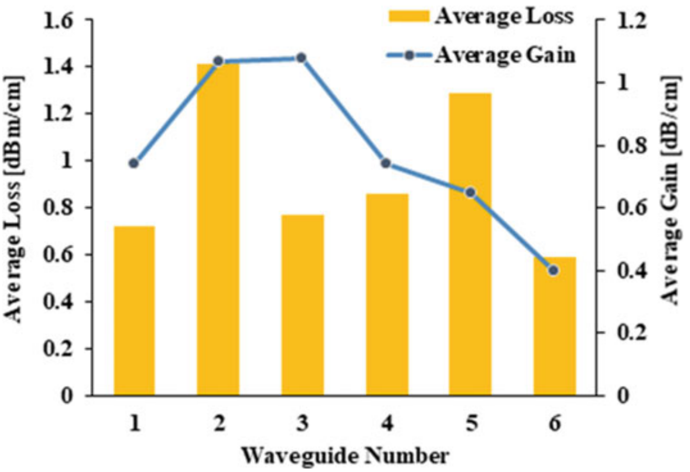


Fig. 4 The average loss and gain for each waveguide number

launching side to couple the −30 dBm input signal of 617 nm red light source and 23 dBm of 532 nm green laser. The data of optical amplification of Eu-Al/BzMA waveguide was then collected by using Optical Spectrum Analyzer (OSA). The collected data are illustrated in Fig. 4.

Based on the result, as expected, the waveguide number that has smaller average core diameter recorded higher average loss as shown by the waveguide number 2 and 5, which is 1.41 and 1.29 dBm/cm, respectively. The satisfactory explanation for higher loss is due to the coupling loss at the launching side, in which the core diameter of the coupler is larger than the size of the respective waveguide core.

For the average gain evaluation, the highest result recorded was 1.08 dB/cm by waveguide number 3 and followed by waveguide number 2, which was 1.07 dB/cm. Even the core diameter for all waveguide numbers are almost the same but, there were variations in the result of optical gain. It is proposed that further analysis and experiments should be conducted on fabrication preparations, along with variations in parameters, such as waveguide length, pump power efficiency, and material concentration have to be varied in order to confirm this situation and the factor that influence the optical gain amplification of Eu–Al/BzMA; RE-M incorporated polymer composite optical waveguide amplifier.

4 Conclusions

Few waveguides with multiple cores of 50 μm circular core graded index (GI) multimode polymer optical waveguides doped with 13 wt% Eu–Al/BzMA were successfully fabricated by using Mosquito method with the precision and efficiency. This study has gone some way towards enhancing the understanding of utilizing the micro-dispenser desktop machine to create the desired core diameter through the Mosquito method. In this research, it was found that a higher average loss could be found at a smaller core diameter, which was 1.41 dBm/cm due to the coupling loss at the launching side. Optical amplification has been observed as high as 1.04 dB/cm with lower insertion loss of 0.77 dBm/cm. Hence, it is believed that the results show the potential in implementing Eu–Al/BzMA; RE-M incorporated polymer composite as a material in an active optical device for visible light communication and short reach network transmission, such as in-vehicle network.

References

1. Tanabe S (2002) Rare-earth-doped glasses for fiber amplifiers in broadband telecommunication. C R Chim 5(12):815–824
2. Soma K, Ishigure T (2013) Fabrication of a graded-index circular-core polymer parallel optical waveguide using a microdispenser for a high-density optical printed circuit board. IEEE J Sel Top Quantum Electron 19(2):3600310
3. Jiang C et al (2002) Fabrication and mechanical behavior of dye-doped polymer optical fiber. J Appl Phys 92(1):4–12
4. Kuzyk M, Paek U, Dirk C (1991) Guest-host polymer fibers for nonlinear optics. Appl Phys Lett 59(8):902–904
5. Miluski P et al (2017) Properties of Eu^{3+} doped poly (methyl methacrylate) optical fiber. Opt Eng 56(2):027106
6. Parola I et al (2017) Fabrication and characterization of polymer optical fibers doped with perylene-derivatives for fluorescent lighting applications. Fibers 5(3):28
7. Milanese D, Schülzgen A, Facciano M (2018) Innovative microstructured optical fibers as waveguides for the infrared region
8. Yoshida Y, Fukui T, Ishigure T (2017) Polymer waveguide incorporated with europium-aluminum polymer composite for compact and high-gain optical amplification devices. In: 2017 conference on lasers and electro-optics Pacific Rim (CLEO-PR). IEEE
9. Mitani M et al (2015) Polymer optical waveguide composed of europium-aluminum-acrylate composite core for compact optical amplifier and laser. In: Integrated optics: devices, materials, and technologies XIX. International Society for Optics and Photonics
10. Mataki H, Fukui T (2005) Organic/inorganic optical nanocomposite with highly-doped rare-earth nanoclusters: novel phosphors for white LEDs. In: 5th IEEE conference on nanotechnology, 2005. IEEE
11. Mataki H et al (2007) High-gain optical amplification of europium-aluminum (Eu^{3+}–Al)-nanocluster-doped planar polymer waveguides. Jpn J Appl Phys 46(1L):L83
12. Mataki H, Fukui T (2006) Blue-green luminescence of terbium-titanium (Tb^{3+}–Ti) nanoclusters. Jpn J Appl Phys 45(4L):L380
13. Tagaya A et al (1995) Theoretical and experimental investigation of rhodamine B-doped polymer optical fiber amplifiers. IEEE J Quantum Electron 31(12):2215–2220
14. Grattan KT, Meggitt BT (1995) Optical fiber sensor technology, vol 1. Springer
15. Yoshida S, Suganuma D, Ishigure T (2014) Photomask-free fabrication of single-mode polymer optical waveguide using the Mosquito method. In: 2014 IEEE photonics conference (IPC). IEEE
16. Yoshida Y, Fukui T, Ishigure T (2017) Polymer waveguide incorporated with europium-aluminum polymer composite for compact and efficient amplification devices. In: 2017 IEEE CPMT symposium Japan (ICSJ). IEEE

17. Kinoshita R, Suganuma D, Ishigure T (2014) Accurate interchannel pitch control in graded-index circular-core polymer parallel optical waveguide using the Mosquito method. Opt Express 22(7):8426–8437
18. Ishigure T (2014) Graded-index core polymer optical waveguide for high-bandwidth-density optical printed circuit boards: fabrication and characterization. In: Optical interconnects XIV. International Society for Optics and Photonics

A Review on Correlating Gas District Cooling (GDC) Model with Data Center (DC) Operation for Environmental and Economic Performance

Nurul Syifa Shafirah Binti Omar, Low Tan Jung and Lukman A. B. Rahim

Abstract This study is a review in finding correlation between supply of chilled water and energy from Gas District Cooling (GDC) model with cooling and energy demand from Data Center (DC) operations. The architecture is called GDC-DC. This architecture was introduced by Hitachi Research in Universiti Teknologi PETRONAS (UTP) as UTP's GDC houses the campus region (academic buildings, chancellor complex and mosque) with electrical energy and chilled water for air conditioners. Based on review from previous research on GDC-DC operations in UTP, the current GDC-DC operations is not meeting the real-time job requirements and energy requirements by DC. Apart from that, DC configurations are inappropriate and non-optimized thus, increasing the power usage of DC and cooling demand. This will contribute to high operational cost on DC and carbon footprint issue due to rise of higher power generation. Eventually affecting the environment and economic performance of GDC-DC. Therefore, this paper aims to find the best real-time scheduling algorithm in DC that will contribute to an optimized DC which will affect the cooling demand. A review is done to help finding the relevant real-time job schedulers which will further to be deployed in the DC model from UTP.

Keywords Gas district cooling (GDC) · Data center (DC) operation · Real-time job scheduling algorithm

N. S. S. B. Omar · Low Tan Jung · L. A. B. Rahim (✉)
Universiti Teknologi PETRONAS, Bandar Seri Iskandar, 32610 Perak, Malaysia
e-mail: lukmanrahim@utp.edu.my

N. S. S. B. Omar
e-mail: nurul_17008738@utp.edu.my

Low Tan Jung
e-mail: lowtanjung@utp.edu.my

Z. Zakaria and R. Ahmad (eds.), *Advances in Electronics Engineering*,
Lecture Notes in Electrical Engineering 619,
https://doi.org/10.1007/978-981-15-1289-6_4

1 Introduction

Data Centers (DC) require high energy consumption and tend to produce large scale of heat energy. In making DC economically and environmentally appealing, cooling model and job distribution algorithm for DC have been proposed. As technology advances, a lot of DCs are trending toward cloud computing. With virtualization, this technology managed to complement the high demand of carrying big data analytics, enterprise resource planning, services, software and many more. Although cloud data center may use less hardware as compared to grid data center, each server in cloud must go through high speed process management. As such, scheduling algorithm for job distribution is introduced. However, the higher the demand for IT services and for data, the better technology is needed. Consequently, there will be more demand of DCs. DCs will therefore be on a very increasing load. More heat will be produced by the servers and more energy will be consumed. As DCs produce more heat, the effect on global warming can be imminent. This will result in higher energy requirements for DC to operate. From the intense usage of electricity, higher carbon footprint in DCs can be a global issue.

One of the direct solutions to counter heat production in DCs is through effective and efficient cooling. Several cooling methods have been introduced by various researchers. Cooling strategies for data center can include three common elements such as heat removal, air distributions and locations of the cooling unit [1]. The cooling method can be broken down into several systems. These are chilled water system, pumped refrigerant for chilled water system, air-cooled system, glycol-cooled system, water-cooled system, air-cooled self-contained system, direct or indirect fresh air evaporative cooling system and self-contained roof-top system. These options can be utilized by DCs based on different cooling process such as server cooling, space cooling, economizer cooling and so on. Conversely, improper set up of cooling systems may result in intense usage of electricity and release greenhouse gas too. Thus, larger carbon footprint. In this research the focus of cooling strategy is towards Gas District Cooling (GDC) in UTP that generates its own electricity to distribute to the whole campus including chilled water production. However, energy requirements of GDC still need to be investigated as the energy requirements of DC operation is dependent to GDC. Therefore, GDC still has its own energy requirements to operate. There are many scheduling algorithms introduced by researchers to enhance the optimization of DC, however, not all algorithm will meet real-time requirements and energy requirements [2, 3]. In this situation, problems can be found when the scheduling algorithm meets the energy requirements, real time requirements will be missed. Apart from that, varieties of jobs distributed in DC may highly impact the DC operations optimization and making the DC having inappropriate or non-optimize configurations. Due to non-optimized configurations, DC is leading to high usage of electricity and increase in power (kWh) usage. Thus, higher cooling demand and more operational cost will effect its economic performance.

As demand for chilled water increases due to higher cooling demand, GDC's chilled water production will increase and resulting in intense of electric usage from Electric Chiller (EC). Therefore, GDC operation is no longer optimized. When GDC & DC both using intense electricity, it will not contribute to reduce carbon footprint, otherwise, it will worsen the impact on environmental performance. The objectives of this research is to find the correlation between real-time job scheduling algorithm and DC machines temperature to balance the chilled water supply-demand gap, at the same time looking forward for potential optimized job scheduling algorithm for DC operation to achieve economical and optimal chilled water demand and to evaluate performance of GDC model with optimized DC operations for better environmental and economical values.

2 Literature Review

In this research, the main objective is to correlate Gas District Cooling (GDC) supply with DC's cooling demand. The location of DC will be HPC3 at Universiti Teknologi PETRONAS (UTP). Cooling demand for DCs in general is relatively high. There are many kinds of method to cool either small scale or large scale of DCs. Cooling aims to remove the heat produced in DCs and to maintain conducive environment for IT equipment to operate and function properly. There are two types of essential cooling. It can be by air or by liquid typically known as chilled water or any kinds of refrigerant. Cooling system for DC involved infrastructures, management and monitoring. So, in this study, the cooling infrastructure will involve supply side from GDC (chilled water supply) and the demand side in DC (cooling demand). Cooling management will include jobs scheduling algorithm in DC. Effective monitoring of temperature in both DC and the supplied chilled water from GDC to identify the optimum temperature for balancing the supply-demand mechanism.

Other than temperature monitoring for supply-demand balancing, another important aspect to be analyzed would be the correlation between GDC supply model and DC demand intensity. That is to be identified via effective and efficient scheduling algorithm to yield the most optimized and energy efficient models for both GDC-DCs linkages. In UTP, GDC plays an important role for cooling most of the buildings in the campus area daily from 7 am to 9 pm daily. GDC has been in operation in UTP since 2005. It was first operated with Steam Absorption Chiller (SAC). However, since 2011, SAC condition has been deteriorating which needs replacement by Electric Chiller (EC) [4]. EC has showed better performance in cooling buildings, but more electricity is used, thus, less 'green' than SAC. Although GDC is not depending on domestic electricity provider since it generates its own electricity, GDC tends to waste more energy instead of saving it because EC consume more electricity than SAC.

The GDC-DC cooling system is supposed to be economical, but due to high demand by DC and the decreasing performance in GDC, an in-depth study to correlate the GDC-DC chilled water supply and cooling demand is vital. So, in this study, simulation of a cluster of cloud data center in HPC3 would be implemented. The demand of cooling energy by DC will be manipulated through several job scheduling algorithms. Then, power usage of each nodes in the cluster will be measured and evaluated. A demand profile is to be generated by this simulation of the cloud DC. To make the correlation realistic, data from GDC in UTP will be collected and analyzed with the demand profile generated. Through this analysis, a new modelling of GDC size and DC operations can be evaluated for optimization. Carbon footprint will thus be estimated and analyzed. Eventually, economic performance can be calculated or determined.

Many researches on DC energy efficiency and cooling demand have been done with several methods. However, GDC-DC architecture is still new. Most of large scale DCs has their own cooling strategy and are isolated from district cooling. In some countries, free cooling can be implemented in DC by using direct cool air and help the chillers consume less electricity [5]. This technique can be applied to countries that has cool air probably during winter or summer nights where the night is much colder than the day. What about countries with hot and humid weather throughout the year?

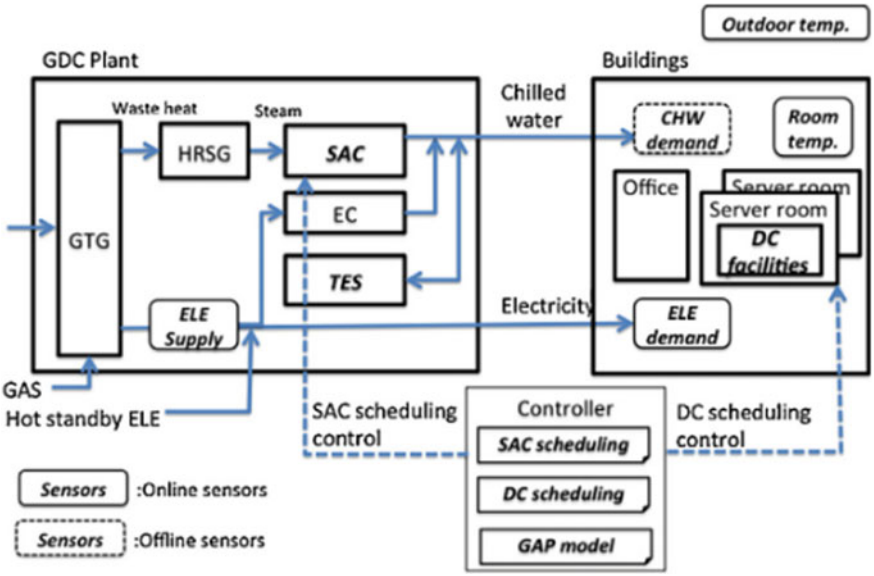


Fig. 1 GDC-DC architecture [9]

Research shows that GDC-DC architecture is probably the suitable method to adequate the cooling and energy demand by DC through supply of energy and chilled water from GDC in ASEAN region with best availability of natural gas resources and pipelines [6]. GDC-DC model in UTP is a product of Hitachi after several research on potential of GDC towards complementing the cooling demand by DC.

Although GDC may help in influencing better environment and economy, due to high demand in utilizing DC, it will produce intense usage of DC and giving high impact in supplying chilled water and demanding of cooling between GDC and DC thus leaving large gap between supply-demand model. Therefore, DC require optimization to maintain its operations and contributing a better performance and influencing better environment and economy like Job scheduling controls in both GDC and DC to have better GDC-DC operations [7–9] (Fig. 1).

2.1 Gas District Cooling (GDC) Model

Gas District Cooling (GDC) provides centralized electrical energy and chilled water to complement the cooling demand throughout campus of UTP. It is equipped with 2 units of gas turbines generators (GTG) where each unit has 4.2 MW capacity, 2 units of heat recovery steam generator (HRSG), 2 units of Steam Absorption Chillers (SAC), 4 units of Electric Chillers (EC), and 1 unit of Thermal Energy Storage (TES) [4]. GDC burns natural gas to produce heat energy to produce electrical energy and the heat then will be recovered using HRSG. Although it produces its own electricity, during normal operation, 2 nos of 4.2 MW generators are operating in island mode, however generators are connected in parallel to domestic electricity supplier, TNB during contingency period [10]. The 2 SACs utilize steam and cooling water to produce chilled water while the 4 EC use electricity to produce chilled water and will be stored in TES [11]. Therefore, GDC still has a slight dependency on TNB (Fig. 2).

The evaluation of system performance of GDC can be determined by calculating the coefficient of performance (COP) of SAC & EC. The performance of GDC should be monitored regularly as GDC operates 24 h daily and the supply of chilled water rely solely on SAC & EC. GDC helps the campus to have less dependency towards domestic electricity supply (TNB), moving towards greener energy generation and greener cooling supply for its district which is UTP. A study has been done indicating that district cooling has less impact on environment. GDC supplies electricity and chilled water with lower CO2 emission than daily electricity-only chilled water supply system [9]. The daily profile of energy consumption for UTP GDC plant to supply chilled water can be explained in Table 1 [9].

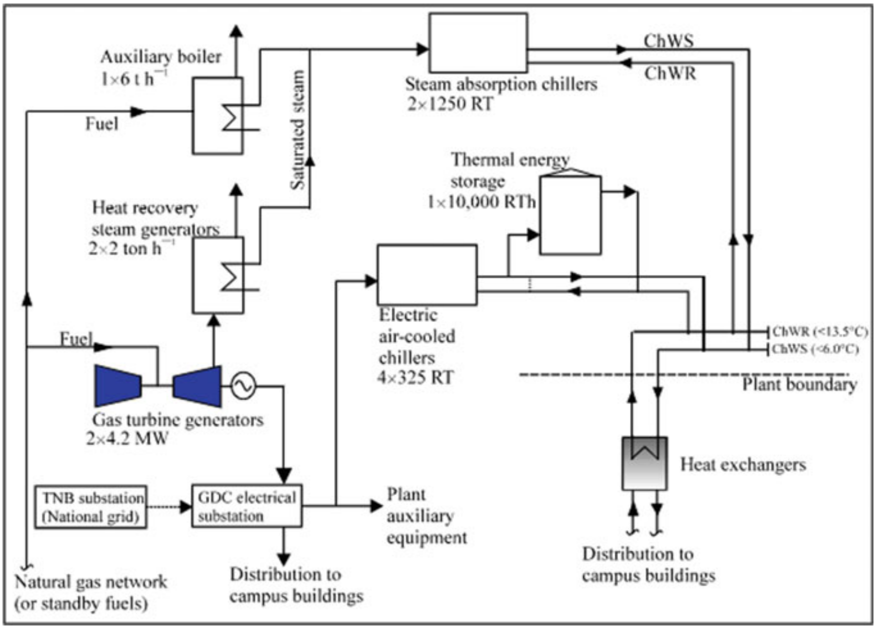

Auxiliary boiler
1×6 t h^{-1}
Fuel
Saturated steam
Steam absorption chillers
2×1250 RT
ChWS
ChWR
Thermal energy storage
1×10,000 RTh
Heat recovery steam generators
2×2 ton h^{-1}
Electric air-cooled chillers
4×325 RT
ChWR (<13.5°C)
ChWS (<6.0°C)
Plant boundary
Fuel
Gas turbine generators
2×4.2 MW
Heat exchangers
TNB substation (National grid)
GDC electrical substation
Plant auxiliary equipment
Distribution to campus buildings
Natural gas network (or standby fuels)
Distribution to campus buildings

Fig. 2 Process flow diagram of GDC plant at UTP [4]

Table 1 Profiling of supply in UTP GDC plant

	SAC	EC	Total
Electricity consumption (MWh)	6.2	11.4	17.6
CO_2 emission [CO_2-ton]	4.2	7.8	12.0
Chilled water supply [MWh]	79.0	61.7	140.7
Chilled water supply per electricity	12.7	5.4	8.0
Chilled water supply per CO_2 [MWh/CO_2-ton]	18.8	7.9	11.7

2.2 Cooling and Optimization of Cloud Data Center

Cloud computing is not completely new concept to be discussed. As workloads increases, demand for higher levels of virtualization, standardization, automation and security is needed. Cloud computing may offer higher performance, capacity and better ease of management. Cloud computing is a specialized distributed computing paradigm is massively scalable, can be encapsulated as an abstract entity where customers outside the Cloud may receive different levels of services, cloud computing is driven by economies of scale, on demand delivery and dynamically configured services usually via virtualization or other methods [12].

On the other side, Armbrust et al. [13] defined that "*Cloud computing refers to both the applications delivered as services over the Internet and the hardware and* *systems software in the data centers that provide those services*." This means, the software and hardware of data center itself is called *cloud*. Armbrust et al. [13] also agreed that cloud computing is available on demand but at the same time eliminate up-front commitment by cloud users, therefore many companies may start small and increase hardware resources as they need from time to time.

As time goes by, the demand for hardware and software resources of data center has increased due to high amount of cloud applications hosting by the data center. As the usage of data center increases, it will contribute a tremendous rise in electricity consumption over the time resulting in high ownership cost of data center and increasing carbon footprints [14, 15]. Therefore, research on energy-efficient and ease-management of data center is a trend nowadays. There are several methods that has been introduced by many researchers that are leading towards energy efficient, reduced electricity consumption, cut down operational cost and lowering emission of carbon footprint of data center operation. The approach and techniques can be found in Table 2.

Table 2 Approach and techniques for data center optimization

No.	Author	Approach/techniques	Objectives/findings
1	Ismaeel et al. [16]	Real-time VM consolidation based on energy minimization and data clustering to reduce error and maintaining low overhead in cloud data center	To forecast cloud data center energy optimization
2	Haruna et al. [17]	Green scheduling algorithm that avoid high thermal stress circumstances while still maintaining the competitive performance in DC through scheduling the jobs to a DC server during minimum room temperature at certain hour	Cooling electricity consumption can be saved with T_aware LSTRF-compared to T_aware FCFS, T_aware RR, and T_aware MLST-RR due different optimization objectives of each schedulers
3	Buyya et al. [18]	Energy-aware data centre resource allocation in VMs migration through modification of Best Fit Decreasing (BDF) algorithm and optimizing current VM resource allocation by selecting VMs according to 4 heuristics	Cloud computing helps in reducing energy consumption cost of data centre
4	Da Costa et al. [19]	Proper cooling infrastructure configuration and power capping techniques. The cooling infrastructure are adjusted to specific types of workloads which results in improvement of CAPEX. Data center operation based on power capping will resulting in addition OPEX savings	CAPEX improvement and OPEX savings

2.3 Real-Time Scheduling Algorithm

In real-time system, the deadlines of tasks can be divided into three different classes, hard, soft and firm real-time system. Real-time system executes a set of tasks according to their temporal constraints. A task consists of several parameters such as execution (*e*), deadline (*d*) and period (*p*). The interpretations of the parameters can be explained when task released a new job in every *p* time units and each job execute for *e* time units and each job must finish before or on *d* time units. Task can be divided into three different categories depending on their arrival pattern which can be explained in Table 3 [20–22].

Real-time scheduling refers to providing a sequence of task execution while considering real-time constraints [23]. Scheduling involvement in execution tasks satisfies the timing constraint [22]. Task in real-time scheduling is said to be feasible, schedulable, optimal predictable, sustainable, utilization bound, preemptive if the job can be interrupted during running state, static or dynamic, and work or non-work conserving [24]. Real time system scheduling can be executed with uniprocessor or multiprocessor. Real-time uniprocessor scheduling gives higher priority to jobs with earlier deadlines, jobs with smaller slack times or least laxity times and jobs with shorter periods or recur at higher rate while real-time multiprocessor scheduling used to solve the meeting of time constraints of real-time task in multiprocessor system. Schedulers for single processor may not be able to apply to multiple processors. Different algorithms utilized to deal with complexity of computational environment.

In real-time multiprocessor system, there are three categories that can be described. These categories of real-time multiprocessor scheduling can be described in Table 4 with a sample of the algorithms [24–27].

Most real-time schedulers should contribute towards optimization and power efficiency of data center. However, each scheduler has its own focus and can be implemented in various types of data center. The schedulers are important in a cluster of distributed computing as it is to enhance the quality of service, to reduce the number of job migrations and to improve resource allocation for the tasks. Introduction to several new resources may increase complexity of schedulers to schedule tasks, although schedulers are said to be optimal but it is only theoretical, yet impractical in the real system [28]. A comparison between environment and functions of data center and types of their optimal schedulers need to be studied as

Table 3 Categories of real-time task

Categories of task	Explanation
Periodic task	The tasks are executed regularly at fixed rates of time
Sporadic task	The tasks are executed and repeated randomly at some bounded rate of time
Aperiodic task	The tasks are executed at any time and has no arrival time pattern thus, no pre-defined timing sequence

Table 4 Categories of real-time multiprocessor scheduling algorithms

RTMS	Description	Scheduling algorithms
Partitioned	Tasks are statically allocated to processors and not allowed to do migration between processors, then each processor is scheduled using uniprocessor scheduling algorithm	First-Fit (FF) Best-Fit (BF) Next-Fit (NF) Worst-Fit (WF)
Global	Allow tasks to migrate between processors and high scheduling overheads is introduced. Shared memory channels are required due to high flow of information	P-fair LLREF EKG LRE-TL DP-Wrap RUN U-EDF
Clustering	A hybrid scheduling of partitioned and global where the processors are grouped, and a scheduler is assigned to each group	DAG

Table 5 Real-time schedulers in LITMUS-RT

Scheduler	Descriptions
P-FP	Partitioned, *fixed-priority*
PSN-EDF	Partitioned, dynamic-priority *earliest-deadline first*
GSN-EDF	Global EDF
C-EDF	Clustered EDF, a hybrid of partitioned and global EDF
PFAIR	Proportionate fair
P-RES	Reservation-based, supports set of *partitioned* uniprocessor reservations—periodic polling server, sporadic polling server, table-driven reservations

some may work with different schedulers out of different environment. In this research, each node in simulation of DC will be running using real-time, Linux testbed called LITMUS-RT. Apart from that, to examine which scheduler provides optimal configuration in data center to balance the supply and demand gap between GDC with DC, 6 real-time scheduling algorithms will be compared as in Table 5.

3 Methodology

Studying correlation between GDC model in UTP with DC operations involves several processes to meet the objectives and to prove that the hypotheses are correct.

In this research, real-time virtualization is expanded in the DC set up. This is to measure whether the current virtualization will meet the real-time requirements or not by running real-time system in our DC. Real-time system can be directed into two types of research. It can be directed using virtualization or multi-core real-time scheduling through hardware partitioning [29]. This paper shows that this research is keen to use hypervisor technology as a direction for virtualization research. Hypervisor is more flexible and efficient due to its ability to share resources and differently emulates the Infrastructure Set Architecture (ISA) within the same physical machine [29].

LITMUS-RT is proposed based on recommendation for experimental platform to perform energy reduction research in real-time system and as an operating system, it allows researchers to explore the behavior of real-time schedulers, thus LITMUS-RT is suitable to be used in this research as a real-time operating system [30]. The real-time system will be running from real-time operating system in each VM (LITMUS-RT) and real-time workloads (benchmarking softwares) will be running in each VM as well. The workloads will be scheduled by credit schedulers in LITMUS-RT. This is to prove whether the real-time schedulers may help the hypervisor to meet the real-time requirements or not. Which will then help to optimize the cloud data center.

Mentioning about optimization, this research focus on the power efficiency of the data center and temperature profile produced by the hardware in cloud data center. Thus, a process of data collection of power consumption and temperature profile while the hosts are running the workloads are currently being obtain through power meter for power consumption and through Linux CPU temperature monitoring software to monitor the temperature of the hosts. Soon, when a set of data has been collected from DC, a further evaluation of environmental and economic performance will be examined and calculated. Based on previous research by Hitachi, chilled water supply-demand gap model can be determined as follows [9]:

$$H_{gan}(t) = |H_{SAC}(t) - H_{demand}(t)| \tag{1}$$

From this model, chilled water supply-demand gap is the difference between SAC chilled water supply and cooling demand. This model is believed to control the supply-demand gap to reduce CO_2 emission. This is because when H_{SAC} is more than H_{demand}, CO_2 emission will increase as SAC is not well used. However, if the H_{SAC} is producing low chilled water less than the H_{demand}, CO_2 emission will still be increasing as the GDC is relatively depending on EC to produce chilled water.

While in DC, the efficiency of power has been measured with several techniques and based on various parameters. There is a metric called Power Usage Effectiveness (PUE) that can be used to determine the energy efficiency of DC. This model is widely used and agreed by Green Grid Consortium [31, 32].

$$PUE == \frac{Datacenter\ Total\ Power\ consumption}{Datacenter\ IT\ Devices\ Power\ consumtion} \tag{2}$$

The facts and figures will help this research to assume the demand for power supply and chilled water from DC that run real-time system which is to serve the users for Infrastructure as A Service (IAAS). Therefore, it can be correlated with GDC side on the supply of power supply and chilled water with their current technology based on their dataset.

4 Results and Discussions

There are six real-time algorithms to be executed in the simulation execution to study their effects on DC operation and influence the correlations at once. Based on literature studies, scheduling algorithms in OS can be analyzed and evaluated based on several criteria such as turnaround time, waiting time, response time, throughput and CPU utilization [33–35]. However, since our DC is implementing real-time system in multiprocessor platform, the algorithms look at Arrival time, T_A, Deadline T_D, Worst case processing time, T_P and Resource requirements T_R [22].

From the literature review, it can be concluded that C-EDF, Clustered-Earliest Deadline First, a hybrid of partitioned and global EDF is more superior than other algorithms in soft real-time case as it generates higher schedulability and achieves lower overheads, while G-EDF or GSN-EDF, Global-Earliest Deadline First, is not preferable for hard-time case as during high-variance utilization scenarios, G-EDF is unable to fully characterize variations. G-EDF is suggested to be implemented in smaller platform or lower processor counts, then P-EDF shows slightly better performance than G-EDF [36]. G-EDF is also less efficient than others with larger overheads due to tick handler of G-EDF which tends to have more lookup of numbers of processor's queue entries in the worst case [37]. Based on Calandrino et al., PD^2PFair and P-EDF tend to perform the best for schedulability with hard-real-time constraints, while in soft-real-time constraints, PD^2PFair and G-EDF performed the best [37].

Further investigation will be done by collecting dataset from power consumption and temperature profile to calculate power usage effectiveness of the small-scale DC operation that will run based on configurations set in methodology. Then, the dataset will be used in quantitative analysis to find out the correlation between GDC model with DC operation to get the balance of supply and demand of energy and chilled water between GDC and DC.

5 Conclusion

Previously, there is no comprehensive correlation study of GDC-DC architecture has been made. Most of the studies only focus on either controlling GDC through scheduling or controlling jobs executions and scheduling algorithm in DC. Then, the performance has been evaluated. This research has a comprehensive view of correlation study on GDC model with DC size to identify the supply and demand relationship of GDC-DC. As for DC operation optimization, from the literature review, it can be expected that PD^{2}PFair will be the best scheduling algorithms to optimize DC operation and compliment the main agenda of this research that is to find correlation between supply of chilled water from GDC with cooling and energy demand from DC in HPC3. From the correlations results, many aspects of environmental and economical can be assessed and referred.

Acknowledgements This project is supported and sponsored by Yayasan Universiti Teknologi PETRONAS (YUTP) grants.

References

1. Evans T (2012) The different technologies for cooling data centers. Schneider Electric White Paper 59, vol 2, pp 1–16
2. Ramamritham K, Stankovic JA, Shiah P-F (1990) Efficient scheduling algorithms for real-time multiprocessor systems. IEEE Trans Parallel Distrib Syst 1(2):184–194
3. Kaur PD, Priya K (2016) Discovering execution of real time tasks in cloud computing. In: Proceedings of the 2015 international conference on green computing and internet of things, ICGCIoT 2015, pp 1048–1053
4. Amear S, Ariffin S, Nordin A, Buyamin N, Amin M, Majid A (2013) Performance analysis of absorption and electric chillers at a gas district cooling plant. Asian J Sci Res 6(2):299–306
5. Dong K, Li P, Huang Z, Su L, Sun Q (2017) Research on free cooling of data centers by using indirect cooling of open cooling tower. Procedia Eng 205:2831–2838
6. Naono K et al (2014) Concept of energy efficient datacenter in ASEAN region. Hitachi Rev 63(9):560–566
7. Okitsu J, Naono K, Sulaiman SA, Zakaria N, Oxley A (2012) Towards greening a campus grid: free cooling during unsociable hours. In: Proceedings of 2012 IEEE conference on control, systems and industrial informatics, ICCSII 2012, pp 202–207
8. Moore J, Chase J, Ranganathan P, Sharma R (2005) Making scheduling 'cool': temperature-aware workload placement in data centers. Science (80-) 61–74
9. Okitsu J, Khamis MFI, Zakaria N, Naono K, Haruna AA (2015) Toward an architecture for integrated gas district cooling with data center control to reduce CO_2 emission. Sustain Comput Inform Syst 6:39–47
10. Bin Khamis MI (2010) Electricity forecasting for small scale power system using fuzzy logic. In: Ipec, 2010, pp 1040–1045
11. Naono K (2014) Performance using linear regression, No. 1
12. Foster I, Zhao Y, Raicu I, Lu S (2008) Cloud computing and grid computing 360-degree compared. In: Grid computing environments workshop, GCE 2008
13. Armbrust M et al (2010) A view of cloud computing. Commun ACM 53(4):50
14. Beloglazov A, Abawajy J, Buyya R (2012) Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. Future Gener Comput Syst 28(5): 755–768
15. Ghribi C, Hadji M, Zeghlache D (2013) Energy efficient VM scheduling for cloud data centers: exact allocation and migration algorithms. In: Proceedings of the 13th IEEE/ACM international symposium on cluster, cloud, and grid computing, CCGrid 2013, pp 671–678
16. Ismaeel S, Miri A, Al-Khazraji A (2016) Energy-consumption clustering in cloud data centre. In: 2016 3rd MEC international conference on big data and smart city, ICBDSC 2016, pp 235–240

17. Haruna AA, Jung LT, Zakaria MN, Haron NS (2016) Green scheduling algorithm for a computational grid environment. In: 2016 3rd international conference on computer and information sciences (ICCOINS)
18. Buyya R, Beloglazov A, Abawajy J (2010) Energy-efficient management of data center resources for cloud computing: a vision, architectural elements, and open challenges. Cloud Computing and Distributed Systems (CLOUDS) Laboratory, Department of Computer Science and Software Engineering, The University of Melbourne, Australia, pp 1–12
19. Da Costa G, Oleksiak A, Piatek W, Salom J (2014) Minimization of costs and energy consumption in a data center by a workload-based capacity management. In: Proceedings of the 3rd international workshop on energy efficient data centers, pp 1–18
20. Shinde V (2017) Comparison of real time task scheduling algorithms. Int J Comput Appl 158(6):37–41
21. Sahoo S, Nawaz S, Mishra SK, Sahoo B (2015) Execution of real time task on cloud environment. In: 2015 annual IEEE India conference, pp 1–5
22. Mohammadi A (2009) Scheduling algorithms for real-time systems
23. Sha L et al (2004) Real time scheduling theory: a historical perspective. Real-time Syst 28(2):101–155
24. Alhussian H, Zakaria N, Patel A (2014) An unfair semi-greedy real-time multiprocessor scheduling algorithm. Comput Electr Eng 50:143–165
25. Davis RI, Burns A (2011) A survey of hard real-time scheduling for multiprocessor systems. ACM Comput Surv 43(4):1–44
26. Carpenter J, Funk S, Holman P, Srinivasan A, Anderson J, Baruah S (2004) A categorization of real-time multiprocessor scheduling problems and algorithms. Handbook on scheduling algorithms, methods and models, pp 1–30
27. Mao H, Schwarzkopf M, Venkatakrishnan SB, Alizadeh M (2018) Learning graph-based cluster scheduling algorithms
28. Lindh F, Otnes T, Wennerström J (2010) Scheduling algorithms for real-time systems. Department of Computer Engineering, Mälardalen University, Sweden
29. Taccari G, Spalazzi L, Claudi A, Broadband A, Taccari L, Fioravanti A (2014) Embedded real-time virtualization: state of the art and research challenges. In: Safe round-trip software engineering for improving the maintainability of legacy software systems (safe RTSE), View project Cluster TAV SHELL_CTN01_00128 (TAV Tecnologie per glia)
30. Borin L, Castro M, Plentz PDM (2017) Towards the use of LITMUS RT as a testbed for multiprocessor scheduling in energy harvesting real-time systems. In: Brazilian Symposium on Computing Systems Engineering, SBESC, vol 2017 – Nov 2017, pp 109–116
31. Liu Z, Wierman A, Chen Y, Razon B, Chen N (2013) Data center demand response: avoiding the coincident peak via workload shifting and local generation. Perform Eval 70(10):770–791
32. Haas J, Froedge J, Pflueger J, Azevedo D (2009) Usage and public reporting guidelines for the green grid’s infrastructure metrics (PUE/DCiE). White Paper, pp 1–15
33. Kishor L, Goyal D (2013) Comparative analysis of various scheduling algorithms. Int. J. Adv. Res. Comput. Eng. Technol. 2(4):1488–1491
34. Dadfar MB, Brachtl M, Ramakrishnan S (2002) A comparison of common processor scheduling algorithms. In: ASEE annual conferences and proceedings
35. Robin R, Putera A, Siahaan U (2016) Comparison analysis of CPU scheduling: FCFS, SJF and Round Robin. Int J Eng Dev Res 4(3):124–131
36. Bastoni A, Brandenburg BB, Anderson JH (2010) An empirical comparison of global, partitioned, and clustered multiprocessor EDF schedulers? In: Proceedings of the real time systems symposium, pp 14–24
37. Calandrino JM et al. (2006) LITMUS^{RT}: a testbed for empirically comparing real-time multiprocessor schedulers, pp 1–5

Multifunctional In-Memory Computation Architecture Using Single-Ended Disturb-Free 6T SRAM

Chua-Chin Wang, Nanang Sulistiyanto, Tsung-Yi Tsai and Yu-Hsuan Chen

Abstract This paper presents an In-Memory Computation (IMC) architecture using Full Swing Gate Diffusion Input (FS-GDI) in a single-ended disturb-free 6T SRAM. Not only are basic boolean functions (AND, NAND, OR, NOR, XOR2, XOR3, XNOR2) fully realized, a Ripple-Carry Adder (RCA) is also realized such that IMC is feasible without ALU (Arithmetic Logic Unit) or CPU. FS-GDI reserves the benefits of the original GDI, and further resolves the reduced voltage swing issue, but it leads to speed degradation and large static power. Therefore, by using in-memory computing technique, the well-known von-Neumann bottleneck will be mitigated as well as energy efficiency is enhanced.

Keywords In-Memory Computing (IMC) · Single-ended 6T SRAM · Ripple-carry adder · Von-Neumann bottleneck · FS-GDI · Boolean functions

1 Introduction

The pursuit of speed in computing system development has never been changed. However, almost all computing architecture used for computation-intensive applications, such as Artificial Intelligence (AI), biological systems, and neural networks, are based on von-Neumann machines, which separates the storage units (memory) with Arithmetic Logic Units (ALU for computation). Thus, despite the advanced CMOS technology, it still runs into a well-known issue called von Neumann bottleneck [1]. Due to the large amount of data flow between memory and CPU and overhead limitations, many types of solutions have been developed, including IMC [2–5]. The aim of IMC is to bypass von Neumann bottleneck and realize computation in memory arrays directly and locally.

C.-C. Wang (✉) · N. Sulistiyanto · T.-Y. Tsai · Y.-H. Chen
National Sun Yat-Sen University, Kaohsiung 80424, Taiwan
e-mail: ccwang@ee.nsysu.edu.tw

Z. Zakaria and R. Ahmad (eds.), *Advances in Electronics Engineering*,
Lecture Notes in Electrical Engineering 619,
https://doi.org/10.1007/978-981-15-1289-6_5

SRAMs, usually as the core of CPU cache, consume a great portion of power. With reference to [6], a 4T load-less SRAM has been proposed and implemented to reduce the power consumption. However, the disturbance of the bit line during read/write data has been pointed out to compromise Static Noise Margin (SNM) [7]. Therefore, a write-assist loop with multi-Vth transistors is presented to ensure the disturb-free feature [8]. Nevertheless, when read/write operations are kept in a long period, the leakage current will destroy the stored data, which needs to be resolved.

Gate Diffusion Input (GDI) technique [9] is a method to relieve two basic problems of Pass Transistor Logic (PTL) circuit. One is the performance degradation from Vth drop, and the other is high power dissipation from half-closed PMOS transistor. Moreover, several boolean functions can easily be expressed by only two transistors. For instance, FS-GDI was revealed to resolve voltage swing hazards [10]. According to the demand mentioned above, a single-ended disturb-free 6T SRAM with IMC architecture utilizing FS-GDI to carry out logic circuit may be a good solution for AI system realizations.

2 SRAM Design with IMC

The proposed single-ended disturb-free 6T SRAM cell with the associated control circuit is shown in Fig. 1. The 6T SRAM cell has been proved to attain the edge of low power and small area. The Control circuit is in charge of generating all the required control signals for the cell. Figure 2 shows an illustrative IMC architecture composed of a 4 × 4 SRAM array, four pre-charged circuits, four MUXs, four RCA unit, and forty-eight 2T switches. Notably, this work also demonstrates a 4-bit Ripple-Carry Adder (RCA) and all the xes in this work (including figures) stand for 0, 1, 2, or 3. Detailed sub-circuits and data flow will be explained below.

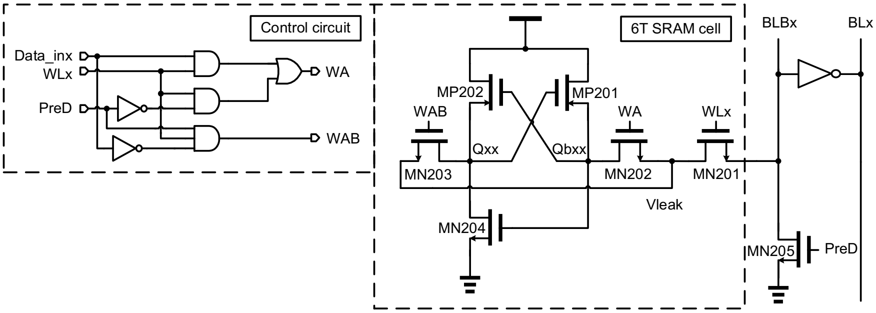


Fig. 1 A 6T SRAM cell with a control circuit. (x = 0, 1, 2, 3)

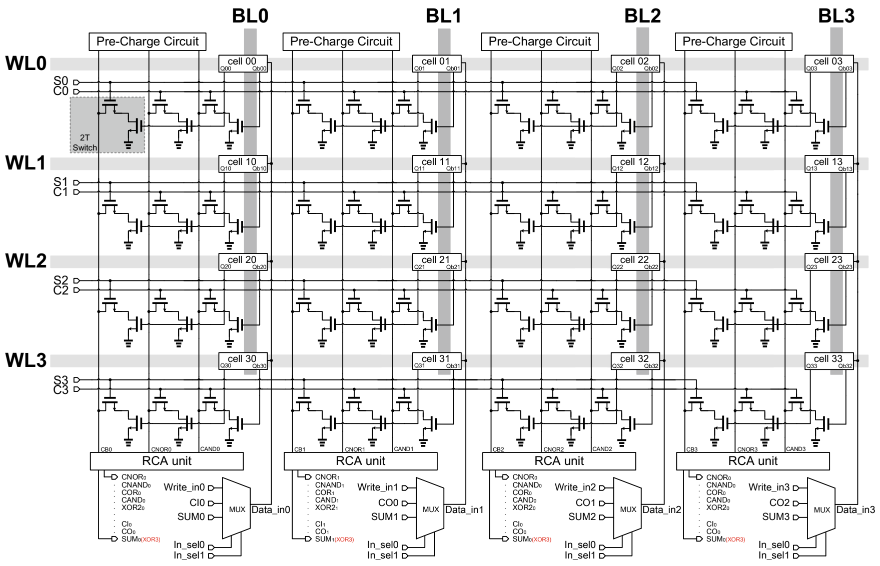


Fig. 2 A 4 × 4 IMC architecture for demonstration

2.1 6T SRAM Cell with Control Circuit

Data_inx in Fig. 1 is the input data to be stored in the cell. PreD is the pre-discharge signal to reset BLBx (BLx). WLx will select which word line to be accessed, and control MN201 to resist the potential disturbance from the bit lines. WA and WAB assist the write operation. If the SRAM cell is realized by the prior 5T SRAM in [8] and Qbxx is logic “1” in read operation, the leakage current will flow through Vleak to Qxx after WA and WLx are switched on. The accumulation on Qxx will soon destroy the data state. Therefore, adding MN204 as a foot switch will fortify the data state on Qxx.

2.2 RCA Unit

RCA unit in Fig. 3 is composed of combinational circuits as well as simplified FS-GDI circuits. Notably, NMOS and PMOS highlighted by grey scale are neglected when one of the inputs are kept coupled to VDD or GND, respectively. Table 1 tabulates detailed logic function in an RCA unit.

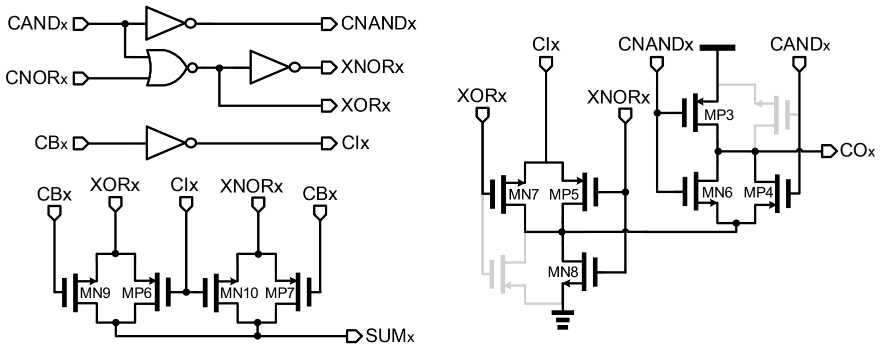


Fig. 3 Combinational logic with simplified FS-GDI. (x = 0, 1, 2, 3)

Table 1 Boolean expressions in the RCA unit

$\mathrm{CBx} = \overline{\mathrm{CIx}}$	$\mathrm{XORx} = \overline{\mathrm{AB} + \overline{(\mathrm{A} + \mathrm{B})}} = \overline{(\mathrm{A} + \mathrm{B})} \cdot (\mathrm{A} + \mathrm{B}) = \mathrm{A} \oplus \mathrm{B}$
$\mathrm{CANDx} = \mathrm{AB}$	$\mathrm{SUMx} = (\mathrm{A} \oplus \mathrm{B}) \oplus \mathrm{CIx} = \mathrm{CIx} \cdot \overline{(\mathrm{A} \oplus \mathrm{B})} + \mathrm{CBx} \cdot (\mathrm{A} \oplus \mathrm{B})$
$\mathrm{CNANDx} = \overline{\mathrm{AB}}$	$\mathrm{COx} = (\mathrm{A} \oplus \mathrm{B}) \cdot \mathrm{CIx} + \mathrm{A} \cdot \mathrm{B}$
$\mathrm{CNORx} = \overline{\mathrm{A} + \mathrm{B}}$	

A, *B* Input bit; *CI* Carry in bit; *CO* Carry out bit

2.3 In-Memory Computing Operation

The IMC operation of the proposed design employs the logic operation strategy reported in [11]. Referring to Fig. 2, the pre-charge circuit will charge CBx, CNORx, and CANDx to high level in the first half of every write cycle. Then 2T switches [12], controlled by Sx and Cx signals (x = 0, 1, 2, 3), will store the digital state in Qxx or Qbxx to CBx, CNORx, and CANDx accordingly. Firstly, only one signal among S0 to S3 will be turned on to read Qxx. If Qxx is high, CBx is low. Secondly, two signals among C0 to C3 will be on to carry out NOR function. If one of the selected Qxx is high, CNORx will turn low. Thirdly, by the same procedure as the previous one, two Cx signals will be on to execute the function of AND gate of Qbxx. If both selected Qbxxes are low, CANDx is high. Overall logic function is tabulated in Table 2. Therefore, input signals, CBx, CNORx, and CANDx, will trigger RCA units to compute the summation and carry bit generation.

For the sake of clarity, we demonstrate X (0101) + Y (0110) = Sum (1011). Logic transition waveforms are shown in Fig. 4. Figure 5 shows the data flow of the 4-bit addition, which is a simplified version of Fig. 2. Notably, the stored data is labeled in red. The data transition of cell blocks is labeled in orange (cell 00, 10, 20, 30, and 21), and the detailed description of the addition is listed below.

Table 2 Logic function table of the RCA unit

S0/C0/C1	Q00	Q10	Qb00	Qb10	CBx	CNORx	CANDx
1/−/−	0	–	–	–	1	–	–
1/−/−	1	–	–	–	0	–	–
−/1/1	0	0	–	–	–	1	–
−/1/1	0	1	–	–	–	0	–
−/1/1	1	0	–	–	–	0	–
−/1/1	1	1	–	–	–	0	–
−/1/1	1	1	0	0	–	–	1
−/1/1	1	0	0	1	–	–	0
−/1/1	0	1	1	0	–	–	0
−/1/1	0	0	1	1	–	–	0

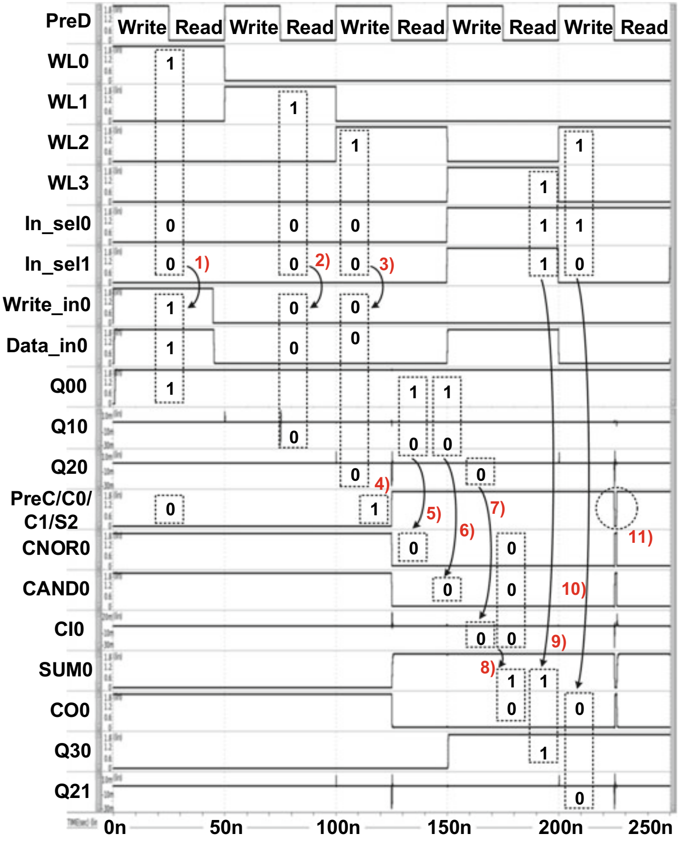


Fig. 4 Detailed logic transitions of an addition

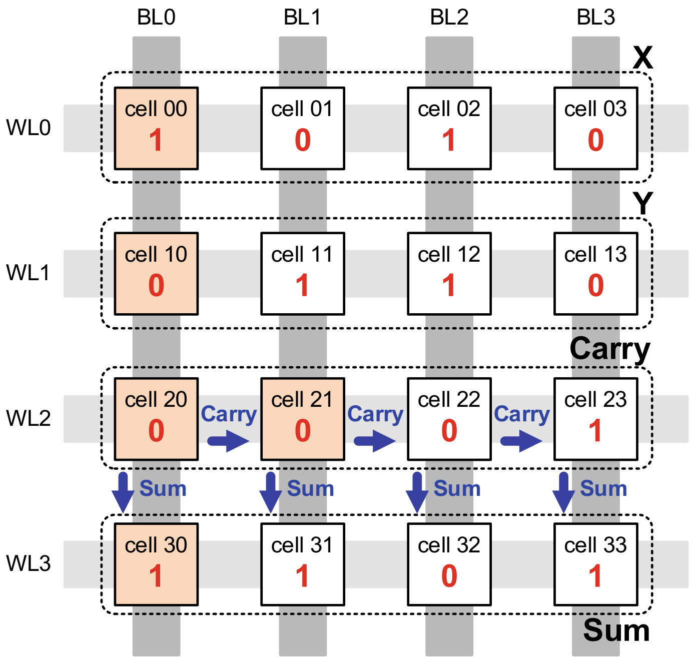


Fig. 5 Demonstration of the operation of the 4-bit ripple carry adder

(1) Enable signal PreC drives the pre-charge circuit to pull CB0, CNOR0, and CAND0 up high (CI0 is low). WL0 is selected to be loaded with data. [0, 0] of In sel[1:0] drives MUX to select Write in 0 as the input data. Q00 is then pulled up high.
(2) Same step as (1). WL1 is then selected to be loaded with data. Q10 is low.
(3) Same step as previous (1) and (2). However, additional carry bit 0 needs to be stored in Q20 to accomplish addition.
(4) PreC pulls high to disable charging, where C0, C1, and S2 are turned on simultaneously to start calculations.
(5) NOR function: Q00 (1), Q10 (0), CNOR0 (0)
(6) AND function: Q00 (1), Q10 (0), CAND0 (0).
7) NOT function: Q20 (0), CB0 (1), CI0 (0).
(8) Addition of bit 0 is complete. SUM0 and CO0 are 1 and 0, respectively.
(9) WL3 is then selected as well as [1] of In sel[1:0] drives MUX to store SUM0 into Q30.
(10) WL2 is finally selected, where [0, 1] of In sel[1:0] enables MUX to reach CO0 as the carry bit to be stored in Q21.
(11) PreC pulls CB1, CNOR1, and CAND1 up high in a short period of time to prepare for the calculation of the next bit.
(12) Repeat steps, (4) to (11) until the calculation is complete.

3 Simulation and Verification

The proposed work is carried out and simulated using UMC 0.18 µm CMOS process. Figure 6 shows the all-PVT-corner simulation (5 Process corners, 3 Voltage variation levels, 3 Temperature) of this 4-bit ripple carry adder. The final result shows that this IMC architecture successfully completes the addition for IMC demand. Comparison with prior IMC SRAMs is tabulated in Table 3. Although we use a legacy CMOS technology in our design, we still attain the least normalized energy on both write and read operations. Most important of all, we are the only ones to realize the addition in a single-ended 6T SRAM.

4 Conclusion

This work presents an IMC ripple carry adder architecture using FS-GDI in a novel single-ended disturb-free 6T SRAM. Not only accumulation problems in original 5T SRAM are resolved, but a simple strategy using FS-GDI to realize the RCA function is proved inside a memory unit.

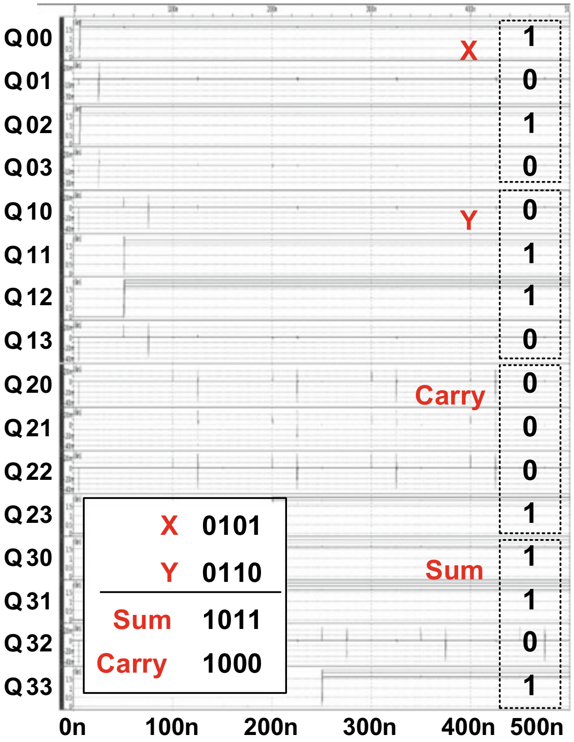


Fig. 6 All-PVT-corner simulation results

Table 3 In-memory computation SRAM performance comparison

	[11]		[12]			*This work*
Year	2018		2018			2019
Process	Fujitsu 55 nm DDC		N/A			UMC 180 nm
Cell type	4 + 2T		8T	8T	8 + T	6T (single-ended)
Operation	AND NOR XOR		NAND NOR XOR RCS	IMP XOR RCS	NAND NOR XOR RCS	NAND NOR XOR SUM
Array size	128 × 128 (16 kB)		N/A			4 × 4
Normalized write energy	28.91 (0.8 V)	88 (0.25 V)	N/A			14.87 (1.98 V, SF, 25 °C) (worst case)
Normalized read energy	25.94 (0.8 V)	78.4 (0.25 V)	N/A			6.99 (1.8 V, TT, 25 °C) (worst case)

$$\text{Normalized write/read energy} = \frac{\text{fJ/bit}}{(\text{supply voltage})^2}$$

Acknowledgements The investigation was partially supported by Ministry of Science and Technology (MOST), Taiwan, under grant MOST 107-2218-E-110-004- and MOST-107-2218-E-110-016-. The authors would like to express their deepest gratefulness to CIC (Chip Implementation Center) in NARL (Nation Applied Research Laboratories), Taiwan, for the assistance of thoughtful chip fabrication.

References

1. Backus J (1978) Can programming be liberated from the von Neumann style? A functional style and its algebra of programs. Commun ACM 21:613–641. https://doi.org/10.1145/359576.359579
2. Wang Y, Yu H, Ni L, Huang G, Yan M, Weng C, Yang W, Zhao J (2015) An energy-efficient nonvolatile in-memory computing architecture for extreme learning machine by domain-wall nanowire devices. IEEE Trans Nanotechnol 14(6):998–1012. https://doi.org/10.1109/TNANO.2015.2447531
3. Jain S, Ranjan A, Roy K, Raghunathan A (2018) Computing in memory with spin-transfer torque magnetic RAM. IEEE Trans Very Large Scale Integr VLSI Syst 26(3):470–483. https://doi.org/10.1109/tvlsi.2017.2776954
4. Jeloka S, Akesh NB, Sylvester D, Blaauw D (2016) A 28 nm configurable memory (TCAM/BCAM/SRAM) using push-rule 6T bit cell enabling logic-in-memory. IEEE J Solid-State Circ 51(4):1009–1021. https://doi.org/10.1109/JSSC.2016.2515510
5. Fan D, Angizi S (2017) Energy efficient in-memory binary deep neural network accelerator with dual-mode SOT-MRAM. In: IEEE international conference on computer design (ICCD). IEEE Press, Boston, pp 609–612. https://doi.org/10.1109/iccd.2017.107
6. Wang C-C, Tseng Y-L, Leo H-Y, Hu R (2004) A 4-Kb 500-MHz 4-T CMOS SRAM using low-V_{THN} bitline drivers and high-V_{THP} latches. IEEE Trans Very Large Scale Integr VLSI Syst 12(9):901–909. https://doi.org/10.1109/tvlsi.2004.833669

7. Wang C-C, Lee C-L, Lin W-J (2007) A 4-Kb low power SRAM design with negative word-line scheme. IEEE Trans Circ Syst I Regul Pap 54(5):1069–1076. https://doi.org/10.1109/tcsi.2006.888767
8. Wang C-C, Hsieh C-L (2016) Disturb-free 5T loadless SRAM cell design with multi-vth transistors using 28 nm CMOS process. In: IEEE international SoC design conference (ISOCC). IEEE Press, Jeju, pp 103–104. https://doi.org/10.1109/isocc.2016.7799754
9. Morgenshtein A, Fish A, Wagner IA (2002) Gate-diffusion input (GDI): a power-efficient method for digital combinatorial circuits. IEEE Trans Very Large Scale Integr VLSI Syst 10 (5):566–581. https://doi.org/10.1109/tvlsi.2002.801578
10. Ahmed MA, Abdelghany MA (2018) Low power 4-bit arithmetic logic unit using full-swing GDI technique. In: International conference on innovative trends in computer engineering (ITCE). IEEE Press, Aswan, pp 193–196. https://doi.org/10.1109/itce.2018.8316623
11. Dong Q, Jeloka S, Saligane M, Kim Y, Kawaminami M, Harada A, Miyoshi S, Yasuda M, Blaauw D, Sylvester D (2018) A 4 + 2T SRAM for searching and in-memory computing with 0.3-V V_{DDmin}. IEEE J Solid-State Circ 53(4):1006–1015. https://doi.org/10.1109/jssc.2017.2776309
12. Agrawal A, Jaiswal A, Lee C, Roy K (2018) X-SRAM: enabling in-memory boolean computations in CMOS static random access memories. IEEE Trans Circ Syst I Regul Pap 65 (12):1–14. https://doi.org/10.1109/tcsi.2018.2848999

Accurate RR-Interval Detection with Daubechies Filtering and Adaptive Thresholding

Mochammad Rif'an, Robert Rieger and Chua-Chin Wang

Abstract QRS detection is needed for electrocardiogram (ECG) signal analysis, including the Heart Rate Variability (HRV) analysis, which is the physiological phenomenon of variation of the time intervals between two consecutive heartbeats. R is the point corresponding to the peak of a QRS complex of ECG waves. RR-interval is defined as the interval between two successive Rs. We proposed an algorithm to acquire RR-interval based on a level-4 Stationary Wavelet Transform (SWT) to decompose ECG signal followed by an adaptive thresholding algorithm to separate QRS complex from other unwanted signals. Daubechies filter is chosen as the mother wavelet, because its shape of the scaling function resembles a QRS complex. The proposed algorithm is simulated by MATLAB, where 48 files from MIT-BIH arrhythmia database are used as benchmarks to verify the algorithm. Simulation results show 99.64% of sensitivity and 99.48% of positive predictivities.

Keywords QRS peaks · RR-interval · Stationary wavelet transform · Daubechies · Adaptive thresholding

1 Introduction

Recently, many people have paid attention to their health, particularly the heart condition. The arrhythmia is one of the heart-related diseases needed to be monitored periodically. It is a problem with the irregular rate of the heart beats. It might be too quick, too slow, or non-periodical. To analyze whether the heart rate is normal or not, it needs to extract features from ECG signals, and the accurate detection of the QRS complex is the critical task in ECG wave analysis, including HRV analysis. HRV is the physiological phenomenon of variation of the time

M. Rif'an · C.-C. Wang (✉)
National Sun Yat-Sen University, Kaohsiung 80424, Taiwan
e-mail: ccwang@ee.nsysu.edu.tw

R. Rieger
Kiel University, Kiel, Germany

Z. Zakaria and R. Ahmad (eds.), *Advances in Electronics Engineering*,
Lecture Notes in Electrical Engineering 619,
https://doi.org/10.1007/978-981-15-1289-6_6

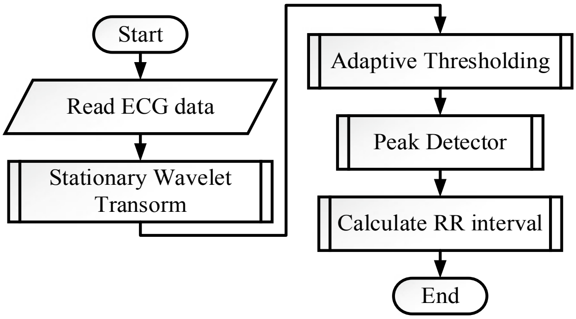
interval between two consecutive heartbeats. Notably, R is the peak of a QRS complex such that RR-interval is defined as the interval between successive Rs. It can indicate many health-related syndromes such as arrhythmia. Many researchers have developed techniques regarding QRS peak detection or RR-interval acquisition. Bayasi, et al. isolated QRS energy centered at 10 Hz with band-pass filtering the raw ECG signal [1]. The filter is composed of low-pass and high-pass filters. Differentiation is then used to find out the high slope such that the QRS complex can be extracted. Zhang, et al. proposed a Pulse-Triggered (PUT) and time-assisted PUT (t-PUT) approach based on the level-crossing events [2]. An event driven by "fall" and "rise" notes is used to detect the QRS complex. Tang, et al. proposed a parallel delta modulator architecture with local maximum point and minimum point detection algorithms to detect QRS and PT waves [3]. A delta modulator represents the slope of the input with a three-state bit stream. Rising, falling, and the difference between rising and falling labels are used as the bit stream.

To resolve the mentioned problems, an RR-interval acquisition using Stationary Wavelet Transform (SWT) followed by an adaptive thresholding algorithm and a peak detector with slope identification is proposed in this study. Thorough simulation with arrhythmia ECG benchmarks is demonstrated to justify the proposed method

2 Accurate RR-Interval Estimation Approach Based on Swt and Adaptive Threshold

Figure 1 shows the flowchart of the proposed method. The discrete ECG signal is decomposed by the SWT to remove noise outside of the desired band. An adaptive thresholding is then applied to separate the QRS wave pattern from the rest. R is the peak of a QRS wave, which needs a peak detector to find out the R-peak point. The proposed method is implemented using MATLAB software and then downloaded to FPGA for hardware verification

Fig. 1 Flow chart of the proposed algorithm for RR-interval estimation



2.1 Wavelet Transform

Theoretically, the signal in time domain or frequency domain can be analyzed by wavelet transforms, particularly, a finite length or fast decaying oscillation signals. The wavelet transform of a signal $f(t)$ is governed by the following Eq. (1):

$$Wf(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} f(t)\psi\left(\frac{t-b}{a}\right)dt \tag{1}$$

where $((t - b)/a)$ is the mother/base wavelet with dilation 'a' and translation 'b'. The higher a, the wider wavelet basis function such that this wavelet coefficient provides low frequency information [4]. For discrete-time signals, the dyadic Discrete Wavelet Transform (DWT) is equivalent, according to Mallat's algorithm, to an octave filter bank [5], which can be implemented as a cascaded of identical cells, e.g., low-pass and high-pass Finite Impulse Response (FIR) filters.

2.2 Stationary Wavelet Transform (SWT)

The typical Discrete Wavelet Transform (DWT) decimates the wavelet coefficients at each level. Thus, the results of the wavelet transform at each level are half the size of the original sequence. The Stationary Wavelet Transform (SWT), on the other hand, pads the corresponding low-pass and high-pass filters with zeros and two new sequences resulting in the generation of the same length as the original sequence [6]. SWT has no decimation in the time domain. By contrast, only a dyadic subsampling of scales (frequency domain) is performed. Hence, it is featured with translation invariance and without resolution loss at lower frequencies, which are major bottlenecks for DWT [6, 7]. Although it increases redundancy in coefficients, the additional artifacts caused by time-domain subsampling at higher scales are avoided.

Based upon the above analysis, SWT with the Daubechies [8, 9] as base/mother wavelet is adopted in this work, because the shapes of the scaling function are close to that of the QRS complex. Low Pass Filter (LPF) for the Daubechies is shown as follows [9]:

$$H_{\varphi}\left(e^{j\omega}\right) = \sqrt{2}\left(\frac{1 + e^{-j\omega}}{2}\right)^{p} R\left(e^{j\omega}\right) \tag{2}$$

where p is vanishing moments [9], and $R(e^{j\omega})$ is a polynomial. The mother wavelet with vanishing moment 3 is chosen because it has a moderate short filter that the computational complexity is low. Notably, the QRS complex power density is positioned in the range of 2–20 Hz, where the maximum is at about 12 Hz [10, 11]. The maximum power spectra of QRS complex is around 8–12 Hz.

Fig. 2 Frequency response of details from db3 SWT

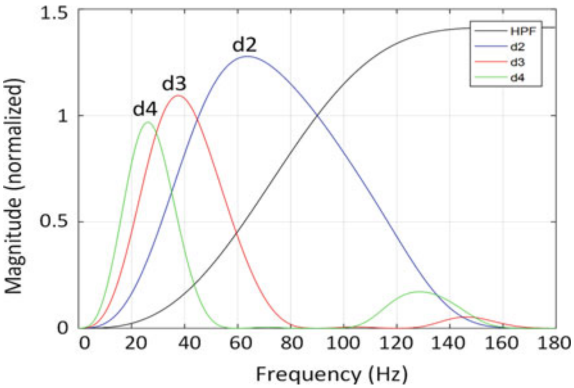


Figure 2 shows the frequency response for several Stationary wavelet transform coefficients with Daubechies as mother wavelet for 360 Hz sampling frequency. The closest bandwidth to the frequency of QRS signals is the level 4 (d4) of SWT with 3 dB bandwidth around 18–34 Hz with the moderate length of filters. Apparently, it is the best choice to enhance the QRS complexes and suppress other unwanted signals or noises, since the R-peak is the only subject for the detection of the interval.

2.3 *Adaptive Thresholding*

The straight forward thought to detect the R-peaks position at level-4 (d4) is to use a threshold. However, the value of the R-peaks at d4 varies along with the time. An adaptive thresholding method is proposed to resolve the time variant issue, where a moving average is calculated as the threshold for the detection of peaks. Figure 3 shows an example of the thresholding, where an average window is sliding along time axis. All d4 coefficients are averaged inside the window. The adaptive threshold value is then generated as the red line, namely the moving average of d4.

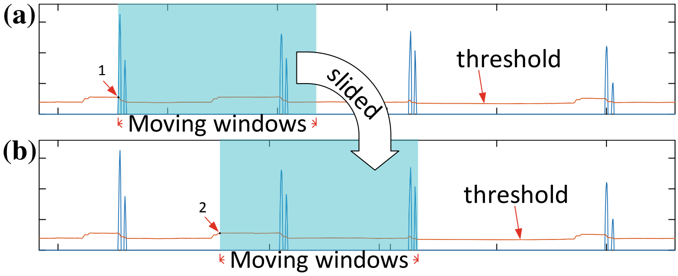


Fig. 3 Moving windows for threshold calculation, **a** at point 1; **b** at point 2

All d4 values below the threshold are reset to zero. By contrast, the data with values over or equal to the threshold are kept the same. Therefore, the d4 processed by the thresholding becomes the blue line in Fig. 3. Notably, the adaptive thresholding is fixed when initialized. This will cause false result for the first beat, because there is no value for d4 at this stage along the look back window.

2.4 Peak Detector

Referring to Fig. 1, the next step of the proposed method is peak detection. Figure 4 shows the peak detector algorithm, where the current data is compared with the previous data such that the slope of these two data is calculated. If the slope is positive, it indicates that the trend is rising, not the peak. By contrast, if the current slope is negative, but the previous slope is positive, the peak is found. The last case is that the current slope and the previous one are both negative.

3 Experiments

A total of 48 records from MIT-BIH arrhythmia database (mtdb) [12] of Physiobank ATM are used for verification. The ECG signals are digitized as 360 samples per second in one channel, where the resolution and voltage range are 11 bits and 10 mV, respectively. All the recordings are annotated independently by cardiologists, which provide reliable and accurate annotation information of each heart beat. Table 1 tabulates the simulation results by the proposed algorithm.

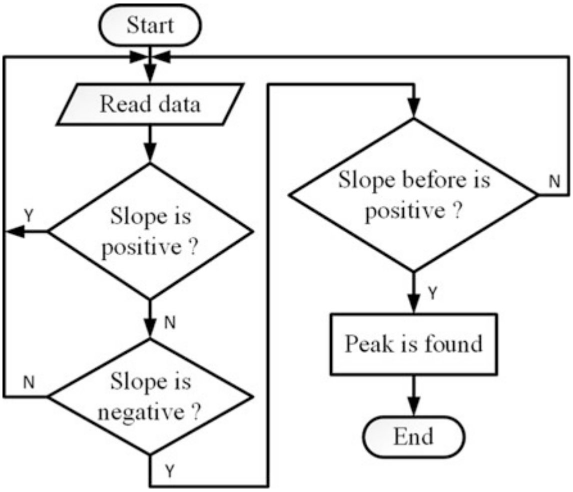


Fig. 4 Peak detector flowchart

Table 1 Summary of performance indexes for 48 records

No	Rec.	FP	FN	TP	Se (%)	P+ (%)	Error (%)
1	100	0	0	2272	100.00	100.00	0.00
2	101	5	0	1864	100.00	99.73	0.27
3	102	0	0	2186	100.00	100.00	0.00
4	103	0	0	2083	100.00	100.00	0.00
5	104	26	0	2228	100.00	98.85	1.17
6	105	22	1	2570	99.96	99.15	0.89
7	106	6	18	2008	99.11	99.70	1.18
8	107	0	0	2136	100.00	100.00	0.00
9	108	36	3	1759	99.83	97.99	2.21
10	109	0	4	2527	99.84	100.00	0.16
11	111	2	1	2122	99.95	99.91	0.14
12	112	1	0	2538	100.00	99.96	0.04
13	113	17	0	1794	100.00	99.06	0.95
14	114	8	0	1878	100.00	99.58	0.43
15	115	0	0	1952	100.00	100.00	0.00
16	116	26	30	2381	98.76	98.92	2.32
17	117	0	0	1534	100.00	100.00	0.00
18	118	1	0	2277	100.00	99.96	0.04
19	119	0	0	1986	100.00	100.00	0.00
20	121	0	1	1861	99.95	100.00	0.05
21	122	0	0	2475	100.00	100.00	0.00
22	123	1	0	1517	100.00	99.93	0.07
23	124	1	0	1618	100.00	99.94	0.06
24	200	9	2	2598	99.92	99.65	0.42
25	201	41	4	1958	99.80	97.95	2.29
26	202	1	3	2132	99.86	99.95	0.19
27	203	13	42	2937	98.59	99.56	1.85
28	205	0	4	2651	99.85	100.00	0.15
29	*207*	*13*	*246*	*2085*	*89.45*	*99.38*	*11.11*
30	208	6	22	2932	99.26	99.80	0.95
31	209	0	0	3004	100.00	100.00	0.00
32	210	4	15	2634	99.43	99.85	0.72
33	212	0	0	2747	100.00	100.00	0.00
34	213	0	1	3249	99.97	100.00	0.03
35	214	2	1	2260	99.96	99.91	0.13
36	215	0	2	3360	99.94	100.00	0.06
37	217	6	5	2202	99.77	99.73	0.50
38	219	1	0	2153	100.00	99.95	0.05
39	220	0	0	2047	100.00	100.00	0.00
40	221	0	3	2423	99.88	100.00	0.12

(continued)

Table 1 (continued)

No	Rec.	FP	FN	TP	Se (%)	P+ (%)	Error (%)
41	222	4	0	2482	100.00	99.84	0.16
42	223	1	0	2604	100.00	99.96	0.04
43	228	44	4	2048	99.81	97.90	2.34
44	230	2	0	2255	100.00	99.91	0.09
45	231	1	0	1570	100.00	99.94	0.06
46	*232*	*215*	*0*	*1779*	*100.00*	*89.22*	*12.09*
47	233	0	5	3073	99.84	100.00	0.16
48	234	0	0	2752	100.00	100.00	0.00
		515	417	109,501	99.64	99.48	0.89

Three indexes, the sensitivity (*Se*), Positive Predictivity (*P*+), and Error are used to compare the performance as follows:

$$Se = \frac{TP}{TP + FN} \tag{3}$$

$$P^{+} = \frac{TP}{TP + FP} \tag{4}$$

$$Error = \frac{FP + FN}{TP + FN + FP} \times 100 \tag{5}$$

The worst result is the false negative of record file 207, as shown in Fig. 5. The output of d4 by SWT can distinguish the QRS complex, though the R peaks are in reverse positions (Fig. 5a) relative to the yellow line Fig. 5b. However, when the noise (pointer 1) and R peaks position (pointer 2) are close, it needs an optimum value for the moving average to tell the difference. The peak detector demonstrates a good result in Fig. 5c to resolve this difficulty.

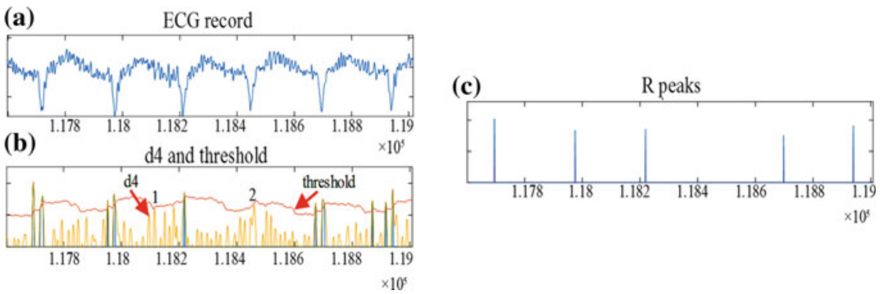


Fig. 5 A part of ECG record file number 207 from MIT-BIH; **a** the original signal; **b** d4 before adaptive thresholding (yellow), after thresholding (blue), and the threshold (red); **c** detected R-peak positions

Notably, the false positive of record file 232 is another issue, where the R-peak values are smaller than the threshold. The noise makes the moving average increase significantly. The key to find the optimum result is the adaptive thresholding, especially the optimum width of the moving window. The average of Se and p+ shown in Table 1 have similar results. It is concluded that the chosen value is near to the optimum one.

To justify the feasibility, the proposed algorithm is implemented and downloaded onto an FPGA (Artic-7). A QRS complex part of record file 100 MIT-BIH arrhythmia as input data (top) is shown in Fig. 6, where clock (clk) period is 100 ns. Peak detection is shown in the bottom strip. There is a delay between two consecutive peaks (R-peaks), which is around several clocks. It is caused by the hardware delay in wavelet implementation. Meanwhile, these peaks at the beginning are not accountable, because the moving average thereof for the dynamic threshold is not reliable.

Performance evaluation of the QRS detector is compared with 3 prior works as shown in Table 2. The values not listed in the reference are written with "n.a.".

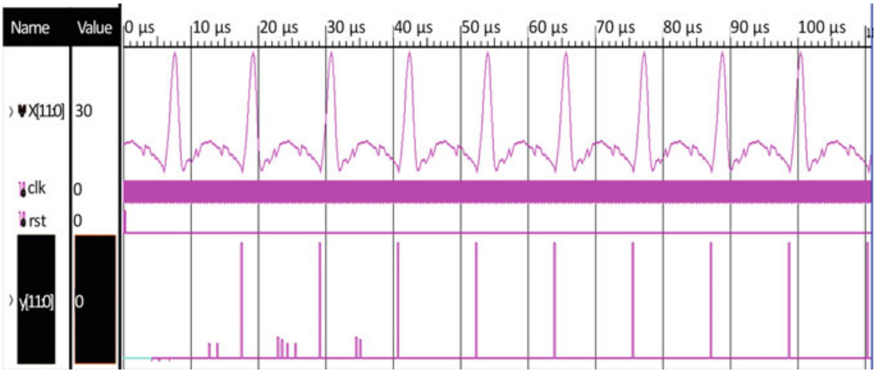


Fig. 6 A part of ECG record file number 100 from MIT-BIH (above), and R-peak position (below)

Table 2 Performance comparison with other method

QRS detector	Technique	TP	FN	FP	Error (%)	Se (%)	P+ (%)
TBCAS [2], 2014	Pulse-triggered	109,966	2665	3015	n.a.	97.63	97.76
ICEE [13], 2017	Level-crossing sampling	n.a.	n.a.	n.a.	1.71	98.89	99.4
TBCAS [3], 2018	Parallel delta modulator	109,966	911	494	1.28	99.17	99.55
This work, 2018	SWT, dynamic threshold	109,501	515	417	0.89	99.64	99.48

Although all of the listed methods have similarity Se and P+, our work demonstrates 0.89% error, which is the best of all.

4 Conclusions

An RR-interval acquisition approach has been presented in this investigation. The algorithm is composed of SWT, moving average for adaptive thresholding, and filtering. 48 records from MIT-BIH arrhythmia database are tested to verify the correctness of the algorithm. 99.64 and 99.48% of sensitivity and positive predictivity for RR-interval estimation are justified, respectively. Last but not least, the proposed algorithm is also successfully implemented by FPGA to prove the feasibility, and the best of 0.89% error

Acknowledgements This research was partially supported by Ministry of Science and Technology under grant MOST 106-2221-E-110-058-, 107-2218-E-110-016-, and 107-2218-E-110-004-.

References

1. Bayasi N, Saleh N, Mohammad B, Ismail M (2014) 65-nm ASIC implementation of QRS detector based on Pan and Tompkins algorithm. In: 10th International conference on innovations in information technology (IIT). Al Ain, United Arab Emirates, 9–11 Nov 2014, pp 84–87
2. Zhang X, Lian Y (2014) A 300-mV 220-nW event-driven ADC with real-time QRS detection for wearable ECG sensors. IEEE Trans Biomed Circ Syst 8(6):834–843
3. Tang X, Hu Q, Tang W (2018) A real-time QRS detection system with PR/RT interval and ST segment measurements for wearable ECG sensors using parallel delta modulators. IEEE Trans Biomed Circ Syst 12(4):751–761
4. Mallat S (2009) A wavelet tour of signal processing. Academic, New York
5. Mallat S (1989) Multifrequency channel decompositions of images and wavelet models. IEEE Trans Acoust Speech Sig Process 37(12):2091–2110
6. Nason G, Silverman B (1995) The stationary wavelet transform and some statistical applications. University of Bristol
7. Merah M, Abdelmalik TA, Larbi BH (2016) R-peaks detection based on stationary wavelet transform. Comput Methods Programs Biomed 121(3):149–160
8. Daubechies I (1988) Orthonormal bases of compactly supported wavelets, communications on pure and applied mathematics, vol XLI, pp 909–996
9. Liu C-L (2010): A tutorial of the wavelet transform
10. Thakor NV, Webster JG, Tompkins W (1983) Optimal QRS detector. Med Biol Eng Comput 21(3):343–350
11. Webster JG (2010) Medical instrumentation application and design. Wiley, New York
12. Physionet. PhysioBank ATM homepage. https://www.physionet.org/cgi-bin/atm/ATM. Last accessed on 23 Sept 2018
13. Ravanshad N, Rezaee-Dehsorkh H (2017) An event-based ECG-monitoring and QRS-detection system based on level-crossing sampling. In: 2017 Iranian conference on electrical engineering (ICEE), Tehran, pp 302–307

Ant Colony Optimization Algorithm to Solve Electrical Cable Routing

W. P. J. Pemarathne and T. G. I. Fernando

Abstract Ant colony algorithms have been applied to solve wide range of difficult combinatorial optimization problems like routing problems, assigning problems, scheduling problems and revealed remarkable solutions. In this paper we are presenting a novel approach of ant colony optimization algorithm to solve the electrical cable routing problem. We have studied the objectives to optimize the cable routing and modified the ant colony system algorithm to get better solutions. This research focused on two objectives, route optimization and the obstacle avoidance. The results of this modified algorithm show an immense improvement in optimizing the electrical circuit designing in a building.

Keywords Ant colony optimization · Electric cable routing · Obstacle avoidance

1 Introduction

Nature offers inspirations to novel solutions for science, technology and engineering. Dorigo and Stutzle introduced the Ant Colony Optimization (ACO) algorithm in 1990s by studying the foraging behavior of real ants [1]. The real ants deposit pheromone while they are walking towards a food source from their nest. With the help of the pheromone density, the rest of the ants can follow the shortest path toward the food sources. Mimicking this behavior into artificial ants has become a popular approach for resolving combinatorial optimization like Traveling Salesman Problem (TSP), quadratic assignment problem and job-shop

W. P. J. Pemarathne (✉) · T. G. I. Fernando
Department of Computer Science, Faculty of Applied Sciences,
University of Sri Jayawardenapura, Nugegoda, Sri Lanka
e-mail: punsi111@gmail.com

T. G. I. Fernando
e-mail: gishantha@dscs.sjp.ac.lk

Z. Zakaria and R. Ahmad (eds.), *Advances in Electronics Engineering*,
Lecture Notes in Electrical Engineering 619,
https://doi.org/10.1007/978-981-15-1289-6_7

scheduling problem [2, 3]. Recently advancements of ACO have been applied to complex problems with multiple objectives to be optimized and proved with remarkable solutions [4, 5].

With the rapid growth in the electrical requirement of the building construction, designing wiring layouts has become complicated. Most of the wiring laying is done manually as a trial and error method. Engineers have presented numerous designs, techniques and systems in order to design the wiring layouts for many types of applications [6]. Optimizing means discovering the best available value of an objective function, or in different types of objective functions. Optimization proves that there are valuable solutions to variables like profit, quality and time [6]. By applying the ant colony optimization to optimize cable routes, it can gain considerable amount of benefits. Optimizing the wiring routes can also help to reduce the length of the wires in the circuit, which will reduce the cost, complexity and the voltage drops of the circuits. Reducing the voltage drops can ensure the safety of the electrical equipment. Automating the wiring routes can also lead to reduction of decision-making time.

The main aim of our research is to modify the ant colony optimization algorithm to optimize the electrical wiring route in order to reduce the length of the wiring circuits and obstacle avoidance. In the recent research literature, we can find few applications of ant colony optimization algorithm in wire and cable routings and harness designing systems. J. M. Alvarado and his team introduced an ant colony systems application for electric distribution network planning [3]. They have used proportional pseudo-random transition rule to explore new paths in order to select the best route. The branches that belong to the best networks found so far were refreshed by applying global pheromone level revision rule. In addition, local pheromone level revision rule is applied, these levels are updated during the route generation process. Their main concern is to minimize the cost and the limits at voltage magnitudes in order to design complete model for the electric system.

Communication cables also need huge attention in building construction. Gianni Di Caro and Marco Dorigo presented a novel approach to adaptive learning of routing tables in wide area best-effort routing in datagram networks [5]. Their study introduced new versions of AntNet, which are AntNet-CL and AntNet-CO. Results shown in both versions had proven better performances compared to the current internet routing algorithm, such as OSPF, SPF, distributed adaptive Bellman-Ford, and recently proposed forms of asynchronous online Bellman-Ford.

Gishantha Thantulage and Tatiana Kalganova have designed a grid-based ant colony algorithm for automatic 3D horse routing [7]. They have applied the algorithm with tessellated format of the obstacles and the generated hoses in order to detect collisions. They have also proved that the algorithm can be applied to real-world hose/pipe routing problems. But when the algorithm is applied to the situations with high resolution or size of the grid, it requires having highest amount of memory and more time to compute the results.

2 Electrical Cable Routing

Cabling plays a key role in building constructions. If one cable goes down the entire system can cease. Therefore, the precise layout design of the entire cable system is very important. Cables designs should be optimized to achieve less complexity, high durability, and low cost. Since the performances of an entire electrical system depends on the cable design, there are few applications to design wire routings in a building. Although, there is no such approach used in the latest nature-inspired algorithms and drawn the results towards optimum solution. Therefore, it is significant to see the practicability and the performance of these nature-inspired algorithms in designing optimized cable layout for a building [8]. Electrical wiring design must follow the recognized standards, use standard electrical symbols and applicable codes. IET (The Institution of Engineering and Technology), which is also known as BS7671, defines the standard regulations for wiring. These regulations focus on the key factors for electrical installations as well as electrical safety [9].

There are considerable number of requirements to be focused on when designing the electrical wiring layout [10]. Mainly, it needs the expert knowledge, and a detail study should be carried out before starting the designing of the circuit. Quality and the requirements, with respect to the price, are also of main considerations. At the studying phase, it is necessary to examine the activities in various types of residences. When laying electrical circuit, there are few constrains to look at architectural design, accessibility, dedicated areas for laying cables, and mounting heights of the electrical accessories must be considered. In this research we have implemented the algorithm based on the radial circuits. Furthermore, the standard mounting heights and dedicated areas for laying cable have been considered according to the BS7671 standards.

3 Ant Colony System

Dorigo and Gambardella [11] derived Ant Colony System (ACS) based on the Ant System (AS) in 1996. They have proven the efficiency of the algorithm by applying it to solve the Travelling Salesman Problem (TSP). ACS is implemented by modifying three main areas of AS.

3.1 An Ant Position on City R Chooses the City S to Move by Applying the State Transition Rule (Pseudo-random-Proportional Rule) Given by Eq. 1

$$s = \begin{cases} \underset{u \in J_k(r)}{\arg\max} \left[[\tau(r,s)] \cdot [\eta(r,s)]^{\beta} \right] & \text{if } q \leq q_0 \text{ (exploitation)} \\ S & \text{otherwise (biased exploration)} \end{cases} \tag{1}$$

where $\tau(r,s)$ is the pheromone density of an edge (r,s), $\eta(r,s)$ is the $[1/\text{distance}(r,s)]$. $J_k(r)$ is the set of cities that remains to be visited by ant k positioned on city r. β is a parameter, which decides the relative importance of pheromone versus distance ($\beta > 0$). q is a random number uniformly distributed in [0, 1], q_0 is a parameter ($0 \leq q_0 \leq 1$), and S is a random variable from the probability distribution given by the Eq. (2).

$$p_k(r,s) = \begin{cases} \frac{[\tau(r,s)]\cdot[\eta(r,s)]^{\beta}}{\sum_{u\in J_k(r)}[\tau(r,u)]\cdot[\eta(r,u)]^{\beta}} & \mathit{if}\ s \in J_k(r) \\ 0 & \mathit{otherwise} \end{cases} \tag{2}$$

3.2 Then the Pheromones Are Updated in the Edges of the Best Ant Tour Using Global Updating Rule in Eq. (3)

$$\tau(r,s) \leftarrow (1-\alpha)\cdot\tau(r,s) + \alpha\cdot\Delta\tau(r,s) \tag{3}$$

where $\Delta\tau(r,s) = \begin{cases} (L_{gb})^{-1} & \mathit{if}\ (r,s) \in \mathit{global\text{-}best\text{-}tour} \\ 0 & \mathit{otherwise} \end{cases}$

$0 < \alpha < 1$ is the pheromone decay parameter and L_{gb} is the length of the globally best tour.

3.3 When Ants Move from One City to Another, Local Pheromone Updating Rule Is Applied as Eq. (4)

$$\tau(r,s) \leftarrow (1-\rho)\cdot\tau(r,s) + \rho\Delta\tau(r,s) \tag{4}$$

where $0 < \rho < 1$ is a parameter. And $\Delta\tau(r,s) = \tau_0$, τ_0 is the initial pheromone level.

4 Algorithm Implementation

The grid points in the wall are designed according to the BS7671 standards and followed the standards of permitted cable routing zones with the mounting heights of the electrical equipment's in dwellings. When mounting switches and socket-outlets for lighting and other equipment in habitable rooms, the appropriate heights between 450 and 1200 mm from finished floor level is considered [10]. This also facilitates the requirements of special assisting people. In addition, the mounting height of wall-mounted socket-outlets and other accessories is required to be enough to keep them suffering from getting wet or impact, which may result from floor cleaning.

Ant completes a circuit by travelling from starting point to the end point, where the socket outlet is located. Initially all the ants are positioned in a starting node, then each ant selects the next grid point to move according to the modified state transition rule in Eq. 5 and using the roulette wheel selection. This will be selected according to the closest and the highest level of pheromone. When an ant builds the tour, local pheromone updating rule in Eq. 4 is applied to the visited points. When the ant reaches to the target point where the power socket is located, then the tour is completed. Then the length of the completed tours of each ant is calculated and only the optimum path is updated with extra amount of pheromone using the global updating rule in Eq. 3.

4.1 Modifications to the Ant Colony System Algorithm to Solve Wire Routing

Modification 1—State Transition Rule

An ant position on city *r,* chooses the city *s* to move by applying the state transition rule (pseudo-random-proportional rule) given by Eq. 5

$$p_k(r,s) = \begin{cases} \frac{[\tau(r,s)]\cdot[\eta(r,s)]^{\beta}}{\sum_{u\in J_k(r)}[\tau(r,u)]\cdot[\eta(r,u)]^{\beta}} & \mathit{if}\ s \in J_k(r) \\ 0 & \mathit{otherwise} \end{cases} \tag{5}$$

where $\tau(r,s)$ is the pheromone density of an edge (r,s),

Heuristic information $\eta(r,s)$ is the [1/distance (s,t)], distance from point s to the target point t.

$J_k(r)$ is the set of cities that remains to be visited by ant k positioned on city r. $\eta(r,u)$ is also taken as [1/distance (s,t)].

β is a parameter, which decides the relative importance of pheromone versus distance (β > 0).

Modification 2—Roulette Wheel Selection

Using the roulette wheel selection, the probabilities are calculated by the state transition rule in Eq. 5 to select the next city to be move, mapped into contiguous segments of a line span within [0 1]. Therefore, each individuals' segment is equally sized to its fitness. A random number is generated and the individual, whose segment spans the random number is selected.

4.2 Pseudo-code of the Ant Colony System Algorithm – Solve Wire Routing

```
Initialization
  turn = 0
Loop
  Release new set of ants from the starting point
  Loop
    turn =turn +1
    For each ant 'a' in the current set
      If ant 'a' does not reach to the target point
        Move to the next point using ACS pseudo random-proportional rule
        (eq. 5) and roulette wheel selection
        Apply local pheromone update rule (eq. 4) to the selected point
      else
        Ant 'a' stops exploring
  Until (reaches the target)
  Apply the global pheromone update rule (eq. 3) to the best path using the ants that reached to the target point
  Remove the current set of ants from the civilization
Until (turn <= MAX_TURNS)
```

5 Experimentation

To analyze the behavior of the new algorithm, three experiments were carried out. Through these experiments, we have applied the algorithm to optimize the distance between start and the end point with obstacle avoidance and then analyzed the behavior of the algorithm in extended grids with multiple obstacles.

The parameter settings for the algorithm were.

The pheromone decay parameter $\rho = 0.1$, $\alpha = 0.1$, $\beta = 2$, $q_0 = 0.9$, $\tau_0 = (n\ L_{nn})^{-1}$ where L_{nn} is the tour length produced by the nearest neighbour heuristic and n is the number of the cities. The heuristic distance $\eta(r, s)$, the [1/distance((s,t))] is the distance from point s to the target point (socket outlet). This will make the solution more effective and feasible. The simulation is conducted on a PC with Intel Core i5-6200U processor (Processor speed = 2.4 GHz, Memory = 8 GB) in the Windows 10 Home operating system using MATLAB (Version R2012b).

5.1 Experiment 1—Wire Routing from Start to a Target Point

The algorithm is initially implemented to the wall design shown in Fig. 1. The grid is designed by adhering to the BS7671 standards height with 15 grid points. In this initial experiment, we have changed the number of ants and the number of the iterations to analyze the behavior of the algorithm to reach towards the optimum length. All the other parameter values are set to the initial values proposed by Dorigo [11] when introducing the ACS algorithm to solve TSP. As shown in Fig. 1 initially, we have improved the algorithm to optimize the distance between start (7.5, 9) and end (4, 0.5) points. We have studied the behavior of the algorithm and recorded the best, average and worst distances and time taken to different combinations of number of ants and the iterations. This has been tested using 25 trials. To analyze the accuracy of the algorithm, we have introduced the same changes to another pair of starting (2.5, 9) and ending (7.5, 0.5) points.

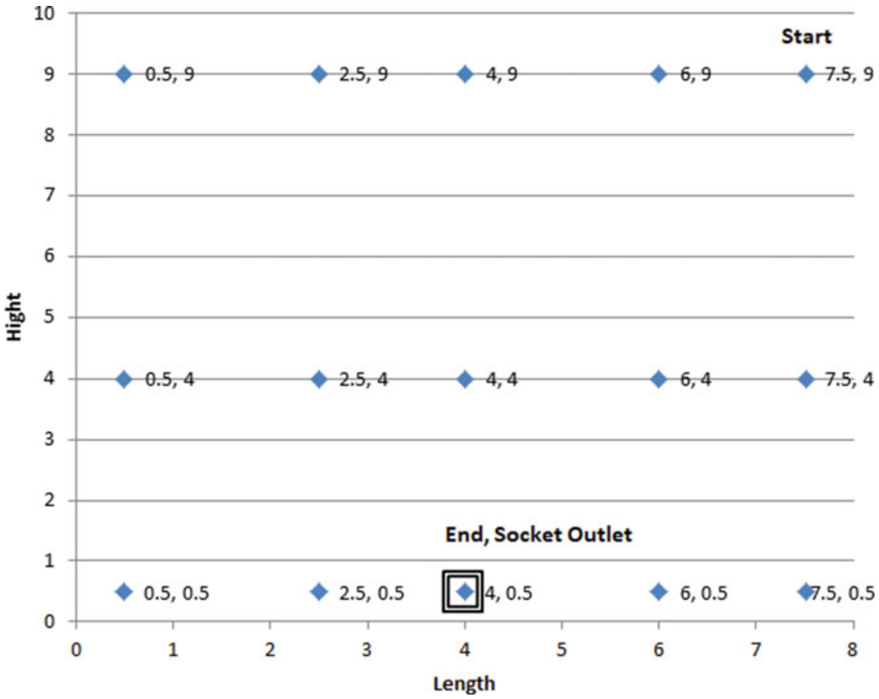


Fig. 1 Model 1–15 point wall design with starting point and the ending point (socket outlet)

5.2 Experiment 2—Wire Routing from Start to Target Point with Obstacle Avoidance

In the next stage, we have introduced an obstacle to this scenario as a door in the wall as shown in Fig. 2. In this scenario, the point behind the door is removed and considered only 11 points around the door. The coordinates of the door are x [5 8 8 5], y [0 7 7 0]. In this scenario, we have introduced a utility matrix shown in Table 1, to avoid the obstacle when selecting the optimum path. Utility matrix was constructed to avoid obstacles and indicate feasible paths ants can move to, if the path is feasible, that is represented as '1', else '0'. Therefore, when the ant is selecting the next point, utility matrix is considered to check if the point is feasible to move or not.

We have studied the behaviour of the algorithm and record the best, average, worst distances and time taken to different combinations of number of ants and the iterations. This has been tested using 25 trials. To analyse the accuracy of the algorithm, we have introduced the same obstacle to the left side of the grid, door is x [0 3 3 0], y [0 7 7 0]. A modified utility matrix is introduced to obstacle

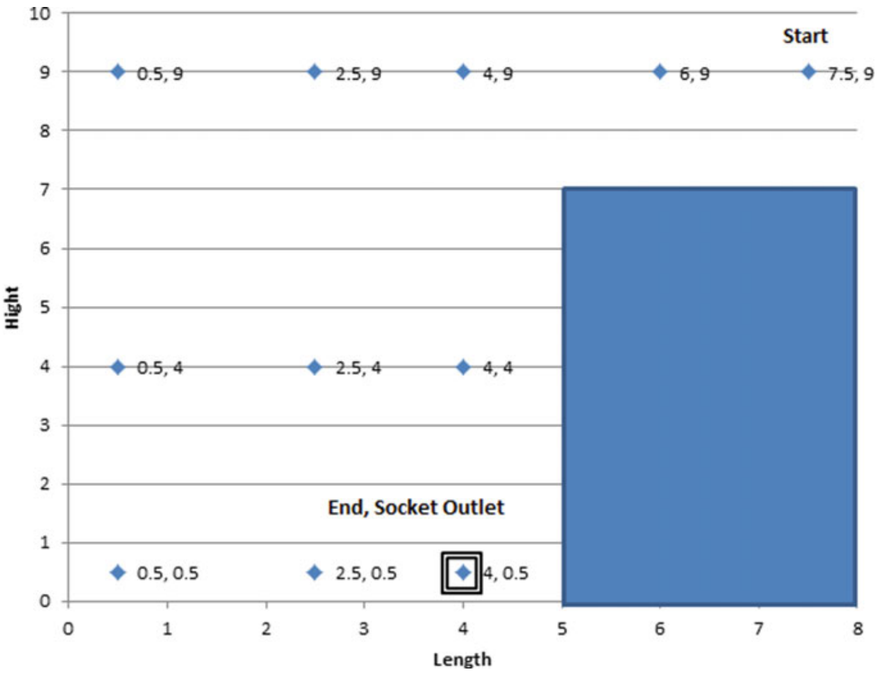


Fig. 2 Model 2–11 point wall design with starting point and the ending point (socket outlet) with a door

Table 1 Utility matrix for model 1

Grid point	1	2	3	4	5	6	7	8	9	10	11
1	0	1	1	1	1	1	1	1	1	1	0
2	1	0	1	1	1	1	1	1	1	0	0
3	1	1	0	1	1	1	1	1	1	0	0
4	1	1	1	0	1	1	1	1	1	1	1
5	1	1	1	1	0	1	1	1	1	1	0
6	1	1	1	1	1	0	1	1	1	0	0
7	1	1	1	1	1	1	0	1	1	1	1
8	1	1	1	1	1	1	1	0	1	1	1
9	1	1	1	1	1	1	1	1	0	1	1
10	1	0	0	1	1	0	1	1	1	0	1
11	0	0	0	1	0	0	1	1	1	1	0

avoidance. In addition, we have introduced another pair of starting (2.5, 9) and ending (7.5, 0.5) points and then recorded the best, average and worst distances and time taken to different combinations of number of ants and the iterations.

5.3 Experiment 3—Wire Routing from Start to Target Point with Extended Grid with Multiple Obstacles

The behaviour of the new algorithm is analysed by introducing to two complex models of expanded grids of more points with multiple objects. As shown in Fig. 3, model 3 consists 44 points and two obstacles. The coordinates of the obstacles are, door x [0 3 3 0], y [0 7 7 0] and the window x [2 12 12 2], y [3 8 8 3]. We have applied the improved algorithm to find the optimized distance between starting point 44 (19.5, 9) and ending point 5 (4.5, 0.5), and then, recorded the best, average and worst distances and time taken through 25 trails. To conduct this experiment, we have considered 20 ants and 1000 iterations.

Model 4 in Fig. 4, consists 48 points and four obstacles. The coordinates of the obstacles are, door x [0 3 3 0], y [0 7 7 0], window1 x [2 4 4 2], y [2 8 8 2], window2 x [7 9 9 7], y [2 8 8 2] and window3 x [12 14 14 12], y [2 8 8 2]. We have applied the improved algorithm to find the optimum distance between starting point 48 (19.5, 9) and ending point 2 (1.5, 0.5) and then recorded the best, average and worst distances and time taken through 25 trails. To conduct this experiment, 20 ants and 1000 iterations were considered.

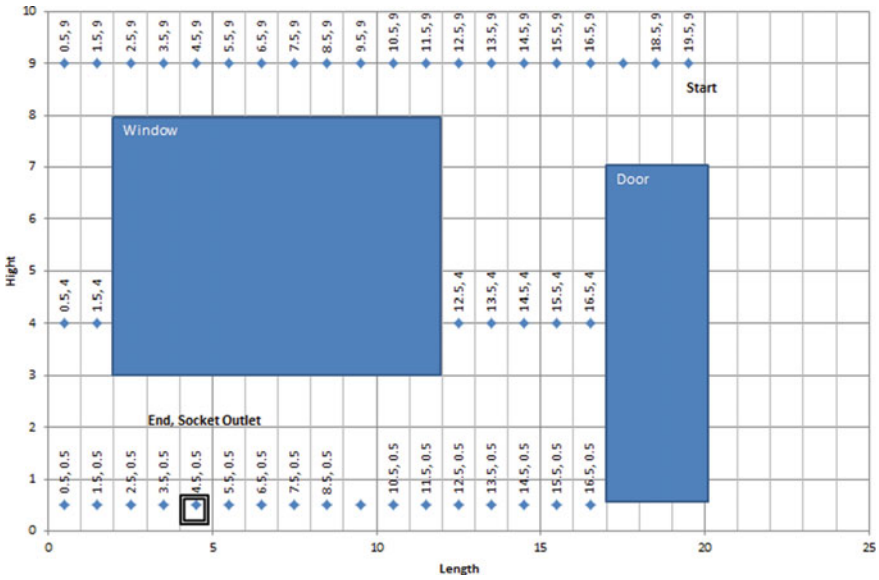


Fig. 3 Model 3–44 points wall design with starting point and the ending point (socket outlet) with 2 obstacles

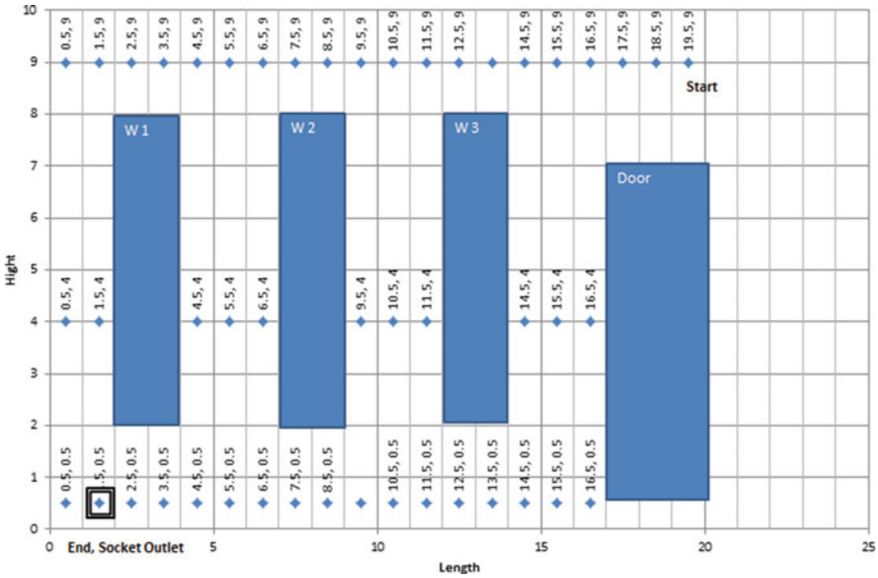


Fig. 4 Model 4–48 points wall design with starting point and the ending point (socket outlet) with 4 obstacles

6 Results and Discussion

6.1 Experiment 1—Wire Routing from Start to a Target Point

As shown in Table 2, experiment 1 is carried out for three different combinations of population of ants and number of iterations to optimize the distance between point 15 and 3. Each combination is tested using 25 trials. Best distance is the direct distance between the points (9.1924 feet) and this is achieved with modified algorithm. 80% of the results proved the direct distance as the best distance. When the population size is 20 and the iterations are 2000, algorithm shows better results, but the time has increased. Since accuracy of the algorithm is also the same for the population size 20 and the iterations were 1000 and the results were produced faster, this is the most suitable combination for the modified algorithm. This has been proved using another set of starting 12 and ending 5 points in the same grid as shown in Table 3. Figures 5 and 6 show the best distance of both instances.

Table 2 Results of experiment 1 model 1—Start point—15, End point—3

Number of ants/ iterations Start point—15, End point—3	Best distance (feet)/time (s)	Average distance (feet)/time (s)	Worst distance (feet)/time (s)
Number of ants—10 Iterations—1000	9.1924/0.8919	10.5418/3.0664	12/6.0115
Number of ants—20 Iterations—1000	9.1924/1.9631	9.6993/5.5841	10.8073/9.5608
Number of ants—20 Iterations—2000	9.1924/9.5649	9.4130/14.4145	10.2321/22.1703

Table 3 Results of experiment 1 model 1—Start point—12, End point—5

Number of ants/ Iterations Start point—12, End point—5	Best distance (feet)/time (s)	Average distance (feet)/time (s)	Worst distance feet)/time (s)
Number of ants—10 Iterations—1000	9.8615/4.5457	11.4485/6.0383	15.1313/9.2218
Number of ants—20 Iterations—1000	9.8615/5.0889	10.8398/6.1229	13.5/6.0094
Number of ants—20 Iterations—2000	9.8615/12.0895	10.02768/14.0554	10.6924/16.5066

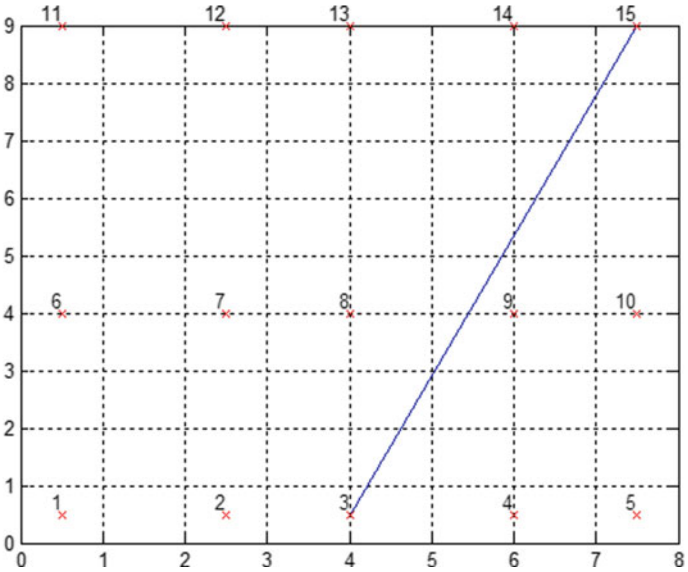


Fig. 5 15 points wall design with starting point 15 and the ending point 3 (socket outlet) output = optimized distance 9.1924

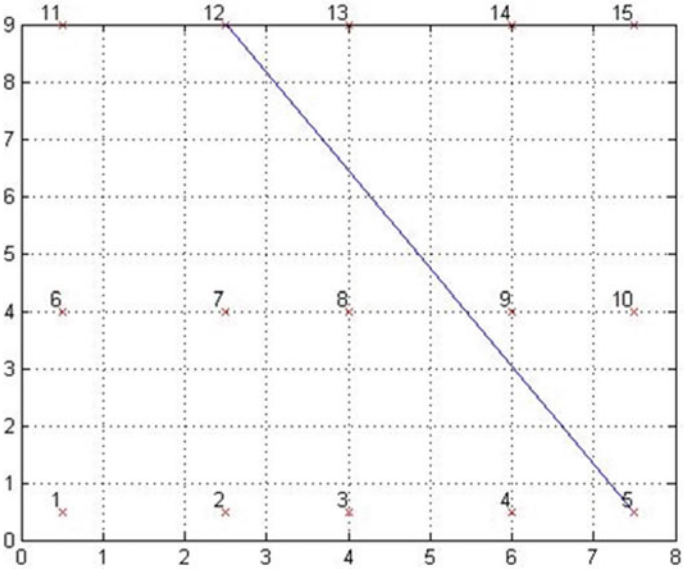


Fig. 6 15 points wall design with starting point 12 and the ending point 5 (socket outlet) output = optimized distance 9.8615

6.2 Experiment 2—Wire Routing from Start to Target Point with Obstacle Avoidance

In this experiment, a single obstacle was introduced to the grid and the modified algorithm was analysed with obstacle avoidance in two scenarios; One with the door in the right side and the other in the left side. As the results show in the Table 4, the best distances given for both scenarios are the same as the shortest distances given by the Dijkstra algorithm. The solution is tested through 20 trials and proved the success rate of 85–90% for both scenarios (Figs. 7 and 8).

Table 4 Results of experiment 2

Model	Best distance (feet's)/ time (s)/accuracy	Average distance (feet's)/time (s)	Worst distance (feet's)/time (s)
Right door—Model 2 Start—11, End—3	11.4112/6.6228 90%	11.5289/7.1989	12/6.9741
Left door—Model 2 Start—7, End—1	10.8852/6.0014 85%	10.9165/6.5522	11.053/5.5213

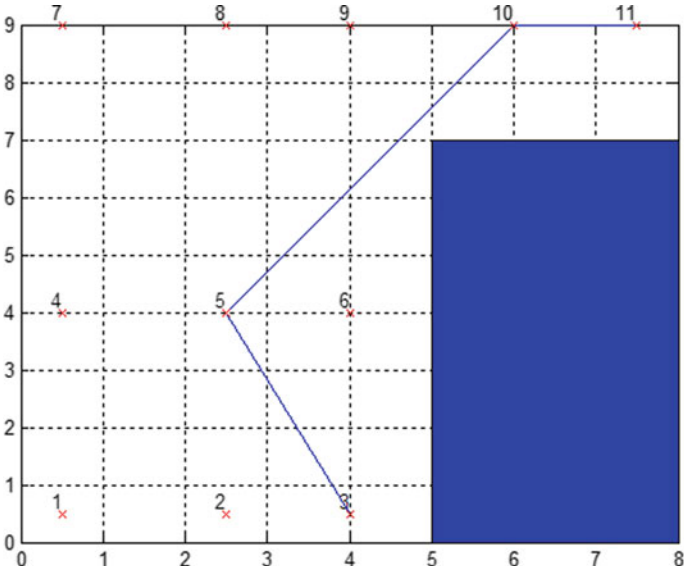


Fig. 7 11 points wall design with starting point 11 and the ending point 3 (socket outlet) with right door. Output = optimized distance 11.4112 feet

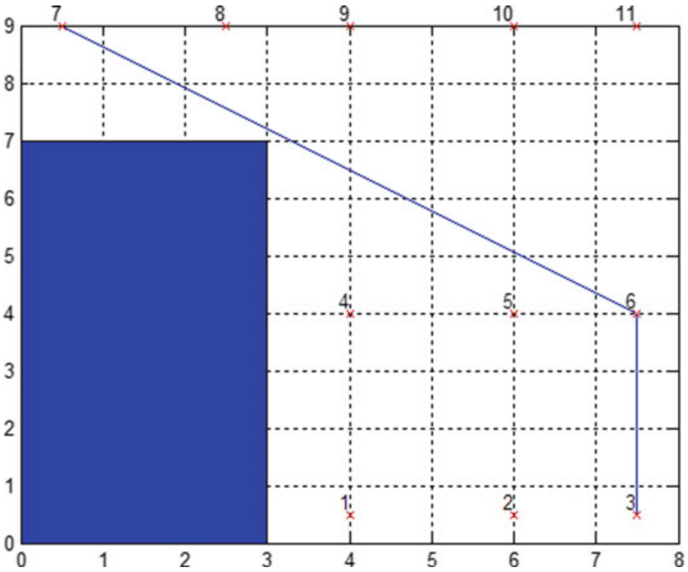


Fig. 8 11 points wall design with starting point 7 and the ending point 3 with left door. Output = optimized distance 9.603 feet

6.3 *Experiment 3—Wire Routing from Start to Target Point with Extended Grid with Multiple Obstacles*

In the third experiment, the modified algorithm was introduced to two grids in model 3 and 4 with few obstacles. In addition, the shortest distances were calculate using the Dijkstra Algorithm, for model 3 from point 44 to 5, the shortest distance is 18.3794 feet and for the model 4, the shortest distance from point 48 to 2 is 20.0139 feet. The modified algorithm also produced the same result as the best distance from same points and it proved 80–85% accuracy from 25 trials as shown in Table 4 (Figs. 9, 10; Table 5).

7 Conclusion

In this study, a new version of ACS has been introduced to generate optimal path between start and the target point. The algorithm is implemented in MATLAB environment. The new algorithm is developed to optimize the length of cables in electrical wiring in a building with obstacle avoidance. Initially, it has considered a 2D single wall with a door. The grid was designed according to the BS7671 standards following the standards of permitted cable routing zones. Results gained through experimenting different models and number of trials were compared again,

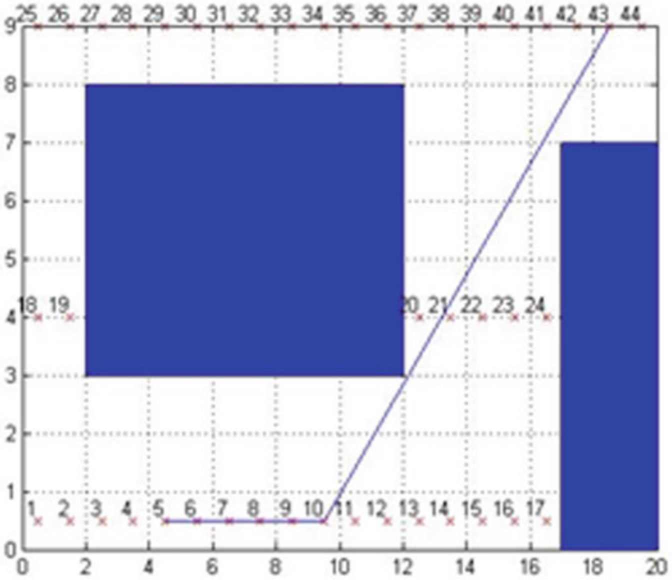


Fig. 9 Model 3–44 points wall design with starting point and the ending point (socket outlet) with 2 obstacles. Optimized distance: 18.3794 feet

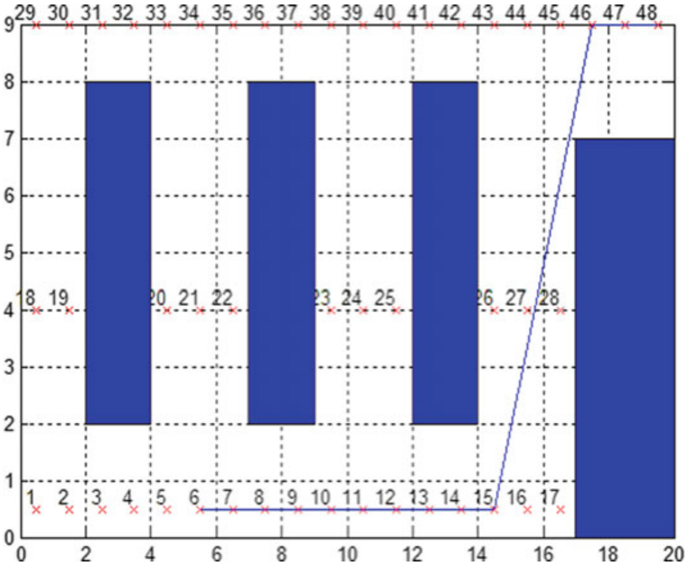


Fig. 10 Model 4–48 points wall design with starting point and the ending point (socket outlet) with 4 obstacles. Optimized distance: 20.0139 feet

Table 5 Results of experiment 3

Model	Best distance (feet's)/ time (s)/accuracy	Average distance (feet's)/time (s)	Worst distance (feet's)/time (s)
Model 3—44 points, 3 objects Start—44, End—5	18.3794/18.679 80%	18.6551/19.3385	19.2424/18.0354
Model 4—48 points, 4 objects Start—48, End—2	20.0139 24.1364 85%	20.38462/24.7678	20.7321/25.24631

for the shortest distances calculated, using a distance matrix. The comparison of results proved that the results given by the modification were the same as the results gained from the Dijkstra algorithm and the modification gave the optimized distance between start and the target point. By applying the above improved ant colony optimization to optimize cable routes, it can gain considerable amount of benefits. Optimizing the wiring routes can help to reduce the length of the wires in the circuit, which will reduce the cost, complexity and the voltage drops of the circuits. Reducing the voltage drops can ensure the safety of the electrical equipment. This also helps to follow electrical wiring standards as well as obstacle avoidance. The results can be finally shown in cable layout map for the engineers to lay circuits efficiently.

The first stage of the research focused on two objectives as optimizing the distance between starting and ending point, as well as avoiding obstacles in the path. At the next stage, the algorithm can be further developed to optimize the path of one starting point to multiple target points. This can be applied to situations in an electrical cabling, where a circuit should cover multiple points. This also can be improved to be used in 3D grid environments. In future, this method can be applied to multi-objectives optimization in order to optimize the number of bends in a circuit with the optimized length, where the graph based deterministic algorithms like Dijkstra and A* are not acceptable.

References

1. Fister I Jr, Yang X-S, Fister I, Brest J, Fister D (2013) A brief review of nature-inspired algorithms for optimization
2. Yang X-S (2010) Nature inspired metaheuristic algorithm, 2nd edn. Luniver Press
3. Alvarado JM, Alvarado EV, Arévalo MA, Quituisaca SP, Gomez JF, Oliveira-De Jesus PM (2009) Ant colony systems application for electric distribution network planning. In: IEEE 15th international conference on intelligent system applications to power systems
4. Warren P (2011) Selecting the right cable system for your environment. W. L. Gore & Associates
5. Di Caro G, Dorigo M (1998) Two ant colony algorithms for best-effort routing in datagram networks. University in the City of Brussels, Belgium

6. Boardman JT, Meckiff CC (1985) A branch and bound formulation of an electricity distribution planning problem. IEEE Trans. Power App. Syst. 104:2112–2118
7. Thantulage GIF (2009) Ant colony optimization based simulation of 3d automatic hose/pipe routing. Ph.D. theses, Brunel University School of Engineering and Design
8. Pemarathne WPJ, Fernando TGI (2016) Wire and cable routings and harness designing systems with AI: a review. In: International conference on information and automation for sustainability (ICIAfS). IEEE
9. Stokes G, Bradle J (2009) A practical guide to the wiring regulations: 17th edition IEE wiring regulations (BS 7671:2008), 4th edn. Wiley
10. Locke D (2008) Guide to the wiring regulations 17th edition IEEE wiring regulations (BS 7671: 2008). Wiley
11. Dorigo M, Gambardella LM (1997) Ant colony system: a cooperative learning approach to the traveling salesman problem. IEEE Trans Evol Comput 1:53–56

Circuit-Based Model Grounding Electrode Considering Frequency Dependent of Soil Resistivity and Relative Permittivity

Ruqayyah Othman, Zulkurnain Abdul-Malek, Muhammad Irfan Jambak, Zainuddin Nawawi and Muhammad Abu Bakar Sidik

Abstract This paper presents the simulation of circuit-based horizontally oriented grounding electrodes with the consideration of frequency dependence of soil resistivity and permittivity. The frequency dependent circuit parameters were determined by the equations given by Sunde and Dwight, while the frequency dependent properties (soil resistivity and relative permittivity) were modeled according to Scott model. Results obtained show that the voltage response is highly affected by the soil resistivity, especially at high soil resistivity value, which in this study, the highest value is 2000 Ωm. This work also considers the front time of the current to study the effect of current front time to the response, and it was found that the faster the current front time, the more significant difference of peak voltage between frequency dependent and frequency independent models is obtained. Obviously, the current circuit-based model (frequency independent model) tends to provide overestimated results. Therefore, the effect of frequency on the soil resistivity and relative permittivity should not be neglected when determining the transient performance of grounding electrode in order to obtain accurate results.

Keywords Circuit-based · Frequency dependence of soil · Resistivity · Relative permittivity · Grounding

R. Othman (✉) · Z. Abdul-Malek
School of Electrical Engineering, Institute of High Voltage and High Current,
Universiti Teknologi Malaysia, Johor Bharu, Malaysia
e-mail: ruqayyah2@live.utm.my

Z. Abdul-Malek
e-mail: zulkurnain@utm.my

M. I. Jambak · Z. Nawawi · M. A. B. Sidik
Department of Electrical Engineering, Faculty of Engineering, Universitas Sriwijaya,
Ogan Ilir, Kota Palembang, Sumatera Selatan 30662, Indonesia

Z. Zakaria and R. Ahmad (eds.), *Advances in Electronics Engineering*,
Lecture Notes in Electrical Engineering 619,
https://doi.org/10.1007/978-981-15-1289-6_8

1 Introduction

Grounding system, which may be consisted as horizontal and vertical electrodes, is an important part in a Lightning Protection System (LPS). When lightning strikes, the high return stroke current disperses into the ground through the grounding electrode, thus, a good grounding system is crucial for human safety and protection of electrical equipment. Many studies and researches had been conducted to improve the overall grounding system performance, and mathematical or theoretical modelling of the grounding system is one of the crucial parts that need to be looked into when trying to improve the transient performance of a grounding electrode. The most popular modelling approaches are known as the circuit-based, transmission line, and electromagnetic field techniques [1–3]. Among these models, circuit model is known to be a simple and the easiest to compute compared to others. However, there is a drawback of this model, which is the accuracy problem.

Previously solved issues on electrode modelling using circuit-based concept are the integration of soil ionization effect, current rate of rise, and frequency dependency of soil under certain conditions [4–11]. The soil ionization effect is covered for all types of grounding electrode, but for frequency dependency of soil resistivity and relative permittivity, there are only two studies conducted on horizontal and vertical grounding electrode under lightning current [12, 13]. The studies, however, only consider the effective frequency (constant frequency value) in the transient analysis instead of taking all the frequencies. Due to that limitation, further improvement can still be made to overcome the accuracy problem of the circuit-based model by considering all the frequencies in the analysis and study the effect of the current front time on the voltage response.

This work aims to develop an improved circuit-based model for grounding electrodes considering the effects of frequency on grounding electrode impedance and voltage response, and it is limited to a single horizontal grounding electrode. The calculation of the circuit parameters is done in MATLAB software and the simulation work is done in CDEGS software.

2 Modeling Methods

2.1 Circuit-Based Model

A lump circuit-based model is used in the transient analysis, where it is only limited to a single horizontal grounding electrode. Figure 1 shows the equivalent lumped circuit consisting R, L, and C, that represents the grounding electrode in Fig. 2 with radius, a, length, *l*, and depth, *d*, buried in the soil. According to the equations given by Dwight and Sunde, the electrode resistances (ohm) of the circuit for horizontal and vertical grounding electrode are given by Eqs. (1) and (2), respectively,

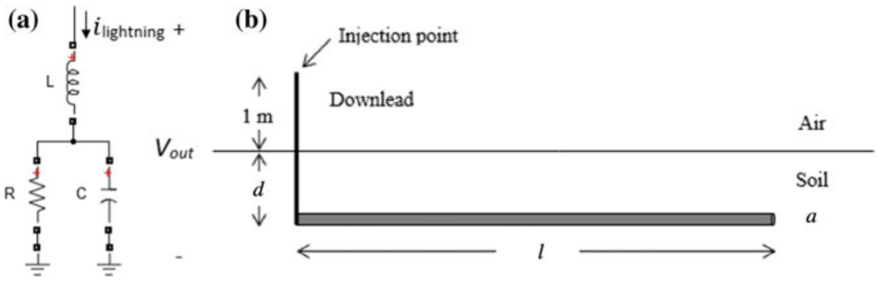


Fig. 1 **a** A lump circuit model of grounding electrode consisting resistor, inductor, and capacitor. **b** Representation of a single horizontal grounding electrode buried in uniform soil, injected with a double exponential current waveform

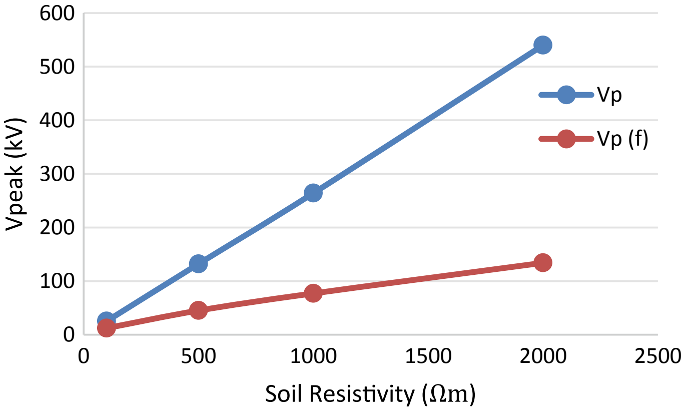


Fig. 2 Peak voltage response at the middle of horizontal electrode by varying the soil resistivity (100, 500, 1000, 2000 Ωm) for frequency dependent and independent models when 10 kA, 1/35-μs impulse injected at one electrode end

$$R = \frac{\rho}{\pi l}\left[\log\left(\frac{2l}{\sqrt{2ad}}\right) - 1\right] \tag{1}$$

$$R = \frac{\rho}{2\pi l}\left[\log\left(\frac{4l}{a}\right) - 1\right] \tag{2}$$

where ρ is the soil resistivity (in ohm-meter), l is the electrode length (in meter), and a is the electrode radius (in meter). The inductance value is given by Eq. (3), which is,

$$L = \frac{\mu l}{\pi}\left[\log\left(\frac{2l}{\sqrt{2ad}}\right) - 1\right] \tag{3}$$

where d is the depth (in meter) of the electrode buried in the soil, and μ is the relative permeability given by $4\pi \times 10^{-7}$. Conductance value can be found by using Eq. (4), which is,

$$C = \frac{\rho\varepsilon}{R} \tag{4}$$

where ε is the soil relative permittivity (in Farad per meter).

2.2 *Frequency Dependence of Soil Parameters*

The proposed frequency dependence of soil model discussed in this section is Scott model, which is used in the analysis due to its accuracy by producing promising results based on the previous studies [12, 14]. The expression of conductivity, $\sigma(f)$, and relative permittivity, $\varepsilon(f)$, as a function of frequency, proposed by Scott are given by (5) and (6), respectively.

$$\begin{aligned}\sigma(f) = {} & 0.028 + 1.098\log_{10}(\sigma_0) - 0.068\log_{10}(f) + 0.036\log_{10}^2(\sigma_0) \\ & - 0.046\log_{10}(f)\log_{10}(\sigma_0) + 0.018\log_{10}^2(f)\end{aligned} \tag{5}$$

$$\begin{aligned}\varepsilon_r(f) = {} & 5.491 + 0.946\log_{10}(\sigma_0) - 1.097\log_{10}(f) + 0.069\log_{10}^2(\sigma_0) \\ & - 0.114\log_{10}(f)\log_{10}(\sigma_0) + 0.067\log_{10}^2(f)\end{aligned} \tag{6}$$

where σ_0 is the conductivity at 100 Hz in (mS/m), f is the frequency in Hertz (Hz) and ε_r is the soil permittivity (F/m).

3 Results and Discussion

In this work, a 15 m long single horizontal grounding electrode, with 0.01 m radius, buried in 1 m depth of uniform soil is considered. The soil relative permittivity is assumed to be constant, which is 1, and the soil resistivity is set to several different values (100, 500, 1000, and 2000 Ωm). The horizontal grounding electrode is injected at a point on top of the download by a double exponential lightning current waveform with 10 kA amplitude, and 1 μs front time as shown in Fig. 1b. The expression of the double exponential current is given by i(t) = 10.244 ($e^{-20000t}$–$e^{-5500000t}$) kA. Based on the results in Fig. 2, it is been shown that the soil resistivity is highly affected by the frequency, especially at high soil resistivity

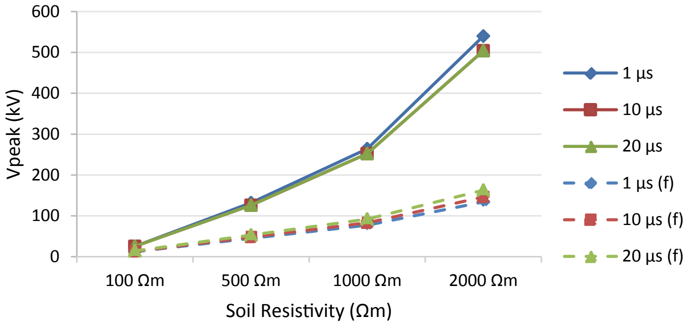


Fig. 3 Effect of input current front times on the electrode voltage for varying soil resistivity for frequency independent (solid-line) and frequency dependent (dash-line) models

value, where the peak voltage of the frequency dependent model is 75.2% lower than the frequency independent model at 2000 Ωm, compared to 52.8% difference at soil resistivity of 100 Ωm.

For the second case, to study the effect of current front time on the response, a double exponential lightning current of 10 kA amplitude with 10 µs and 20 µs front time are considered. The results obtained are compared to observe the effect on the voltage response as shown in Fig. 3. It is found that the response is highly affected by the frequency at fast current front time (1 µs), where the peak voltage of the frequency dependent model is 17.6% lower than the peak voltage at slower front time (20 µs) at soil resistivity of 2000 Ωm compared to 8.2% at 10 µs front time. If the response is being compared to the frequency independent model, the fast front time peak voltage gives 75.2% difference compared to 67.7% difference at slow front time. It is noticed that the grounding electrode voltage response is highly affected by the frequency at fast current front time and high soil resistivity value.

4 Conclusion

The circuit-based model of grounding electrodes with frequency dependent soil was successfully modelled. The values of soil resistivity and permittivity decrease with the applied frequency. Consequently, this has caused the grounding electrode voltage response to be significantly affected (up to 70% decrease in grounding electrode voltage), especially for soil with high resistivity (1 kΩ m and above). Fast current front time (1 µs front time in this case) has found to be significantly affecting the voltage response as the frequency is considered (up to 18% lower in grounding electrode voltage compared to slow current front time of 20 µs). Thus, it can be concluded that, instead of assuming constant values, the effects of frequency on the soil resistivity and permittivity need to be considered when determining the transient performance of a grounding electrode.

Acknowledgements The authors wish to thank the Ministry of Higher Education (MOHE) and Universiti Teknologi Malaysia (Research Vote Nos. 4F828, 4B345, 4B379 and 18H10) for the financial aid.

References

1. Grcev L (2009) Time-and frequency-dependent lightning surge characteristics of grounding electrodes. IEEE Trans Power Delivery 24(4):2186–2196
2. Grcev L (2009) Modeling of grounding electrodes under lightning currents. IEEE Trans Electromagn Compat 51(3):559–571
3. Trlep M, Jesenik M, Hamler A (2012) Transient calculation of electromagnetic field for grounding system based on consideration of displacement current. IEEE Trans Magn 48 (2):207–210
4. Mokhtari M, Abdul-Malek Z, Salam Z (2015) An improved circuit-based model of a grounding electrode by considering the current rate of rise and soil ionization factors. IEEE Trans Power Deliv 30(1):211–219
5. Araneo R, Maccioni M, Lauria S, Geri A, Gatta F, Celozzi S (2015) Hybrid and pi-circuit approaches for grounding system lightning response. In: 2015 IEEE Eindhoven PowerTech. IEEE, The Netherlands, pp 1–6
6. Yutthagowith P (2016) A modified pi-shaped circuit-based model of grounding electrodes. In: 2016 33rd international conference on lightning protection (ICLP). IEEE, Portugal, pp 1–4
7. Yutthagowith P, Kunakorn A, Potivejkul S, Chaisiri P (2012) Transient equivalent circuit of a horizontal grounding electrode. In: 2012 International conference on high voltage engineering and application. IEEE, China, pp 157–161
8. Mokhtari M, Abdul-Malek Z, Gharehpetian GB (2016) A critical review on soil ionisation modelling for grounding electrodes. Arch Electr Eng 65(3):449–461
9. Mokhtari M, Abdul-Malek Z, Salam Z (2016) The effect of soil ionization on transient grounding electrode resistance in non-homogeneous soil conditions. Int Trans Electr Energy Syst 26(7):1462–1475
10. Mokhtari M, Gharehpetian GB (2018) Integration of energy balance of soil ionization in CIGRE grounding electrode resistance model. IEEE Trans Electromagn Compat 60(2): 402–413
11. Mokhatri M, Abdul-Malek Z (2014) The effect of grounding electrode parameters on soil ionization and transient grounding resistance using electromagnetic field approach. Trans Tech Publ Appl Mech Mater 554:628–632
12. Mokhtari M, Abdul-Malek Z, Wooi CL (2016) Integration of frequency dependent soil electrical properties in grounding electrode circuit model. Int J Electr Comput Eng 6(2):792
13. Othman R, Abdul-Malek Z (2018) An improved circuit-based grounding electrode considering frequency dependence of soil parameters. In: 2018 International conference on electrical engineering and computer science (ICECOS). IEEE, Indonesia, pp 271–274
14. Cavka D, Mora N, Rachidi F (2014) A comparison of frequency-dependent soil models: application to the analysis of grounding systems. IEEE Trans Electromagn Compat 56 (1):177–187

A Comprehensive Study on Deep Image Classification with Small Datasets

Gayani Chandrarathne, Kokul Thanikasalam and Amalka Pinidiyaarachchi

Abstract Convolutional Neural Networks (CNNs) showed state-of-the-art accuracy in image classification on large-scale image datasets. However, CNNs show considerable poor performance in classifying tiny data since their large number of parameters over-fit the training data. We investigate the classification characteristics of CNNs on tiny data, which are important for many practical applications. This study analyzes the performance of CNNs for direct and transfer learning based training approaches. Evaluation is performed on two publicly available benchmark datasets. Our study shows the accuracy change when altering the DCNN depth in direct training to indicate the optimal depth for direct training. Further, fine-tuning source and target network with lower learning rate gives higher accuracy for tiny image classification.

Keywords Deep image classification · CNN · Transfer learning

1 Introduction

Image classification [1–3] is one of the major tasks in computer vision investigated for many years. Many application areas, such as image captioning [4], object tracking [5, 6], scene understanding [7], and event detection [8], for a multitude of other purposes [9, 10], used image classifying as the primary task. Compared to

G. Chandrarathne · A. Pinidiyaarachchi
Department of Statistics and Computer Science,
University of Peradeniya, Peradeniya, Sri Lanka
e-mail: gayani.indunil@gmail.com

A. Pinidiyaarachchi
e-mail: ajp@pdn.ac.lk

K. Thanikasalam (✉)
Department of Physical Science, Vavuniya Campus,
University of Jaffna, Jaffna, Sri Lanka
e-mail: kokul@mail.vau.jfn.ac.lk

Z. Zakaria and R. Ahmad (eds.), *Advances in Electronics Engineering*,
Lecture Notes in Electrical Engineering 619,
https://doi.org/10.1007/978-981-15-1289-6_9

many approaches [3, 11], machine learning shows the most promising method to classify images in human accuracy [12]. With the advancement of deep learning, the recently introduced Deep Convolutional Neural Networks (DCNN) showed state-of-the-art performance in image classification [1, 12, 13]. These DCNNs are able to outweigh challenging problems, background clutter, deformation, occlusion, and variations in viewpoint and scale in image classification. The gain of these DCNNs is that they come together with feature extractor and classifier, which used to be two separate processes, such as Support Vector Machine (SVM) [2], k-nearest neighbour [11], logistic regression, and decision trees for classification and, HOG [14], SIFG [15], as hand-crafted feature extraction methods used with those classifiers. Further, the DCNNs are capable of binary or multi-class classifications, in which a set of images is classified into one label or set of labels.

Obtaining significant results with deep learning is very difficult since deep learning approaches are often required enormous amount of images. It is shown that if a huge amount of data is available, CNNs with large number of layers are capable of achieving better than human performance in visual recognition [12]. The excellent results were obtained from DCNNs trained on large scale image classification dataset [16], and also have a large number of classes. The DNN requires large data sets in order to generalize the model properly and to avoid over fitting. When analysing the remarkable DNN models, which showed state-of-the-art results, it is clear that they contain a large number of layers in order to distinguish the large number of classes very effectively. These deeper networks facilitate a good learning capacity performance exceptionally with large datasets, since the larger number of layers is able to handle the large amount of parameters in the dataset. Even though it is convincing that DNNs require large datasets, identifying the most suitable size of the dataset is not possible and there is not any indication of the relation between the dataset size and the deep learning model. Hence, this situation becomes more tragedic since most of the real time applications suffer due to the data limitations [17]. For the time being, researchers arbitrary choose the deep learning architectures to start training model for classifications.

On the other hand, there is no ‘fit for all’ minimum dataset size to train a DNN. But, training deep CNN model with less number of layers or training with small datasets prevents getting more accurate results from the model due to over fitting and under fitting problems. CNN models with few layers are unable to utilize the hierarchical features of a larger dataset through those limited layers and cause the lower accuracies. Therefore, training a DCNN with small datasets is hectic because a small dataset in a DNN leads to under fitting the model. Due to these parameter limitations, CNNs are still struggling to reach the state-of-the-art accuracy for tiny data applications. Image classifier with limited data is beneficial for many real world application areas such as object tracking, scene understanding, and real-time since many areas struggle with data dependency of deep learning. Collecting labelled data sets are very expensive in many areas, such as medical images, environmental science, etc. [18]. Collecting larger datasets are challenging due to the expensive data labelling and in some applications, images are limited.

To overcome the data limitations, researchers are trying to increase the data using different techniques. Data augmentation [1] is a common technique to increase the training data by generating data virtually. Even though these techniques increase the data by generating additional images, CNN models still cannot overturn the issue. In this work we analyse the performance of CNN based image classification with tiny datasets. This paper aims to investigate deep learning approaches in order to obtain better results with small datasets. We compare direct training results, by changing the depth of the network. Then we test the transfer learning approach to see the accuracy increment compared to the direct training. We selected two publicly available benchmark datasets for this study. The rest of the paper is organized as follows: Section two describes the background including the related work. Problem overview, methodology, results, and discussions are stated in sections three to six, respectively. Finally, the paper is concluded in section seven.

2 Background

2.1 Image Classification with CNNs

CNNs are a special type of Neural Networks that work in the same way of a regular neural network, except that they have neurons in three dimensions. Although CNNs were introduced two decades back [19], a major breakthrough in image classification with CNN was the model called as AlexNet proposed by Krizhevsky et al. [1] (shown in Fig. 1). After the great success of AlexNet, various CNN architectures were proposed for image classification, such as VGG-Net [13], GoogleNet [20], and ResNet [12]. Most of the later introduced CNNs for image classification are mostly based on these popular architectures and they improve the performance by increasing the number of convolutional layers [13, 20]. Thus, typically, CNN for classification tasks is composed with a set of convolution layers and fully connected layers stacked on top of each other as the convolution layers at the beginning and fully connected layers on top of those convolutional layers. The convolutional

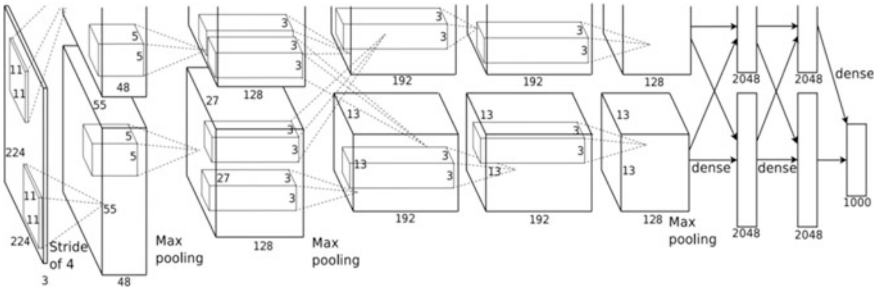


Fig. 1 AlexNet Architecture. It has eight layers: five convolutional and three fully connected

layers of these CNNs extract the features and then the fully-connected layers perform the classification task [13, 21, 22]. As shown in Fig. 1, the popular AlexNet architecture has five convolutional layers and three fully connected layers. Convolution layers represent image features in a layered hierarchical manner [21].

The depth of CNNs or the number of layers in a CNN perform a crucial work in a better classification model. By varying the depth of the CNN, the learning capacity of CNNs can be controlled [23]. By investigating the introduced models in ILSVRC, it can be seen that if the model is deepened [1, 12, 13, 20], the accuracy of the model is increased. But, the model depth only can be increased until the model accuracy is saturated [12]. After that, typical stacked layered CNN architectures are not capable to minimize the error [12]. As alternatives to the typical stacked layered architectures, GoogleNet proposed a hierarchical convolutional layer structure for classification, and ResNet proposed deeper network (up to 152 layers) architecture with a less complex structure.

The internal learning structure of CNN architectures has been investigated for various applications [21, 22]. It clearly shows that CNNs are representing the image features in a layered hierarchical structure. CNN layers tend to learn generic features, such as edges, intensity, and colour information in the first few layers. Therefore, it is necessary to understand that more application specific features, are learned by last few convolutional layers. Figure 2 shows the feature representation of popular VGG-16 architecture. As shown in the figure, while first convolutional layers represent more generic (low-level) features, last layers represent more specific (high-level) features.

Ultimately, the important fact is identifying a good architecture with the concerns of model depth and the learning patterns. The selected CNN architecture should have the appropriate learning capacity for the available dataset and the capability of generalizing the dataset without over fitting. But there is no method of identifying the number of layers based on the dataset. Consequently, many attempts have been taken to classify small datasets using CNN [25, 26]. Even though these

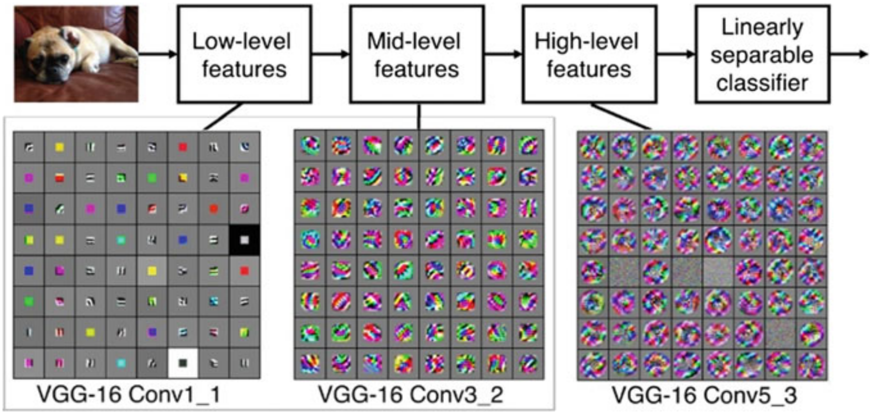


Fig. 2 Visualizing layers of VGG-16 [13] architecture. Image taken from [24]

approaches improve the performance considerably, they face the over-fitting problem because of direct training (from scratch). So it leaves the question whether shallow architectures are capable enough to capture all features for small data or not. If not, how to train a large CNN architecture with tiny data. Therefore, recent approaches avoid direct training and improve the performance by transfer learning approaches and some other techniques.

2.2 Deep Transfer Learning

Very recently, CNN models started using transfer learning approaches to training data [27]. These approaches have addressed the data limitation issue by transferring learned knowledge from one application domain to another relevant domain [27]. This transfer learning technique provides a better way to avoid training CNNs from scratch.

Fundamental motivation for transfer learning for machine learning discussion started in 1995 with NISP workshop on learning to learn [28], which was focused on finding methods to retrain and reuse previously learned knowledge. Nowadays, this technique is widely used in image classification tasks [17, 29]. As shown in Fig. 3, The objective of this technique is to transfer the knowledge from a source model to a target model [18, 27]. Furthermore, to train the target model with a little amount of target data by transferring learned knowledge from a source model. The hypothesis of transfer learning is that the initial layers of source network are able to represent generic features, hence, can be transferred to other tasks. The similarity between source and target data decides [18] the level of knowledge that can be transferred [18]. If source and target data are much similar, knowledge can be transferred from high number of source layers. If dissimilar, fewer numbers of layers can be transferred. There are studies that are conducted to analyse their performance and measure the level of transferring knowledge [18].

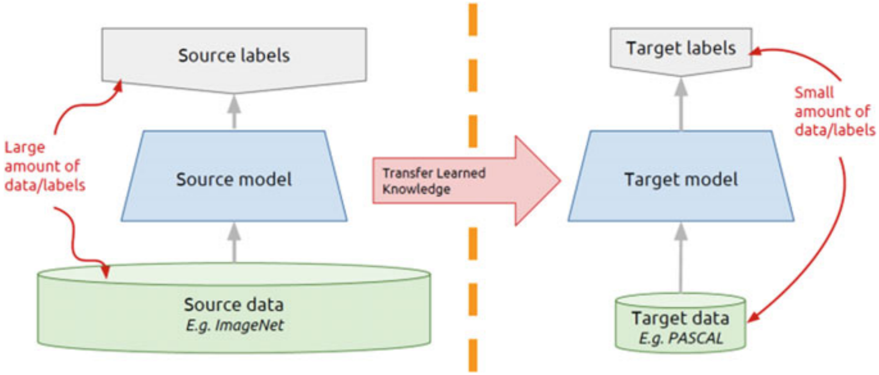


Fig. 3 Deep transfer learning diagram

In practice, knowledge is transferred from pre-trained CNNs, which are trained on large scale datasets, such as ImageNet [16]. Also, some researchers used the pre-trained CNN on a different task as a generic feature encoder and trained a shallow classifier (such as SVM) by using these deep convolutional features [30]. Razavian et al. showed that their "Off-the-shelf" approach outperforms traditional classification approaches with a large margin [31]. As some other knowledge transfer approaches, Ganin and Lempitsky [32] proposed an unsupervised domain adaptation technique and Kokul et al. [6] proposed an online domain adaptation technique for visual tracking by fine-tuning VGG-M network. Deep domain adaptation (transfer learning) approaches [33, 34] are widely used for classifying small data, which are some other attempts taken with transfer learning. These approaches are classifying small data by transferring the learned knowledge from a source task to a similar task.

3 Problem Overview

3.1 Problem Specification

As noted above, it is challenging to train a complex model, which uses only a small amount of training data. On the other hand, this model is unable to specify the suitable architecture or the number of layers that needs to be on a CNN model. The objective of this work is to conduct a comprehensive study on deep image classification techniques for tiny data and to analyze their performance on publically available benchmark datasets. As an initial step, we train CNN architecture directly for tiny data and measure the performance. Then we conduct the deep transfer learning and measure the performance at different levels of translation. In addition, we fine-tune the CNN architecture and measure the performance. We have selected two publicly available benchmark datasets for this study.

3.2 Datasets

We selected CIFAR10 [35] and Caltech101 [36] as tiny image datasets. ImageNet [16] dataset is selected as source data for deep transfer learning. Details of these datasets are given in Table 1.

Table 1 Dataset Comparison

Dataset	Caltech101	CIFAR10	ImageNet
#Classes	101	10	1000
Image size	200 × 300[a]	32 × 32	469 × 387[a]
Images per class	40–800	10,000	120,000[a]

[a]indicates average values

As shown in Table 1, Caltech101 and CIFAR10 (target datasets) have smaller numbers of images per classes compared to ImageNet (source dataset). Even though most class categories are the same in target data and source data (such as bird, airplane and cat), image size, resolution and alignment are different. Test set of CIFAR10 contains exactly 1000 randomly-selected images from each class. We used 20% of randomly-selected images of each class to test the Caltech101.

4 Methodology

4.1 Training from Scratch

As the first step, we train CNN architecture from scratch. We design the network based on VGG-M architecture. Our network takes input with the size of 32 × 32 and produces output as the number of classes in the dataset. In our architecture, convolutional layers are followed by ReLU activation [1]. Max pooling layers are placed in between convolutional layers. Convolutional layers are followed by three fully connected layers with the neurons of 512,256, number of classes (E.g. 10 for CIFAR10). Fully connected layers are interleaved with ReLU activation layer and dropout [37].

We conduct the scratch training evaluation process as follows:

1. The number of fully connected layers kept fixed.
2. The number of convolutional layers increased from small amount to high.
3. The whole network is trained with lower learning rate.

Classification accuracies are measured for the corresponding number of convolutional layers.

The high performing CNN architecture for CIFAR10 is shown in Fig. 4.

We conducted the training for each dataset through 100 epochs with the learning rate of 0.001. We have used Stochastic Gradient Descent (SGD) to optimize the training. The performance of the network is measured as average classification accuracy of the test set. The classification accuracy is measured as follows:

$$Classification\,Accuracy = \frac{Number\,of\,True\,Object\,Detections}{Total\,Number\,of\,Test\,Images} \tag{1}$$

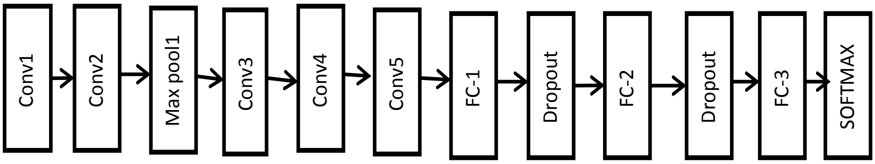


Fig. 4 Proposed CNN architecture for image classification from scratch

4.2 Training from Deep Transfer Learning

In the second step of this study, we conduct the analysis of deep transfer learning based training. We have selected ImageNet and VGG-16 as the source data and the source network, respectively. Figure 5 shows the convolutional layer structure of VGG-16 architecture. We add three full connected layers with the convolutional layer structure of VGG-16. The number of neurons in fully connected layers is set as 512, 256 and the number of classes of each dataset. Fully connected layers are interleaved with ReLU activation and dropout. The Stochastic Gradient Descent (SGD) is used to optimize the network.

In this step, we do not change the number of convolutional and fully connected layers. To measure the performance of transfer learning, we conduct the training process as follows:

1. The convolutional layers of VGG-16 are Re-initialized from top to bottom.
2. The re-initialized layers are trained (fine-tuned) with lower learning rate.
3. Remaining convolutional layers are frozen (kept fixed) throughout the training.
4. Classification accuracies are measured for corresponding number of fine-tunned layers.

We fine-tune the network through 100 epochs with the learning rate of 0.001. The network is trained with the batch size of 32 and SGD is used to optimize the training. We follow the Eq. 1 to measure the classification accuracies.

5 Results

5.1 Implementation Details

This study is implemented in python with Keras [33]. This evaluation is conducted on four cores of 2.66 Intel Xeon with NVIDIA Tesla K40 GPU. The pre-trained VGG-16 model is obtained from Keras model zoo.

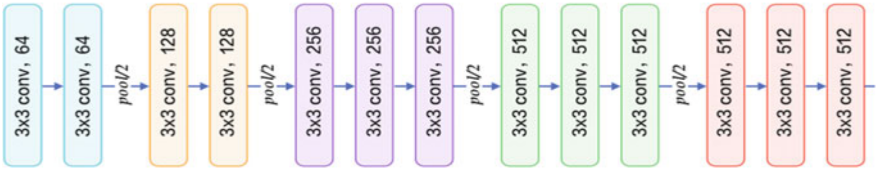


Fig. 5 CNN layer structure of VGG-16 architecture. Kernel size and number of neurons are denoted in each layer

5.2 Scratch Training Results

We started the scratch training with a base network, which had two convolutional layers and three fully connected layers. Then the numbers of convolutional layers were increased one by one and for each architecture, corresponding classification accuracies were measured.

Figure 6 demonstrates the training and testing accuracies of scratch training for the CIFAR10 dataset. As can be seen from the graph, test accuracy is increased smoothly with the number of convolutional layers until the architecture reaches five layers. The maximum test and train accuracies are obtained at five-layer architecture with the values of 90 and 82.5, respectively. The accuracies start falling from six layer architecture. In addition, the difference between training and testing accuracies is increased while we include more convolutional layers.

The scratch training results of Caltech101 is shown in Fig. 7. The accuracy variation of this dataset is most similar to the results of CIFAR10. While comparing with CIFAR10, Caltech101 achieves the higher classification accuracy with fewer numbers of layers. The four convolutional layer architecture gives high training and test accuracies with the value of 86.8 and 79.6, respectively.

5.3 Transfer Learning Results

Figure 8 shows the classification accuracies of transfer learning for both datasets. We have used the VGG-16 architecture as the base network to evaluate the

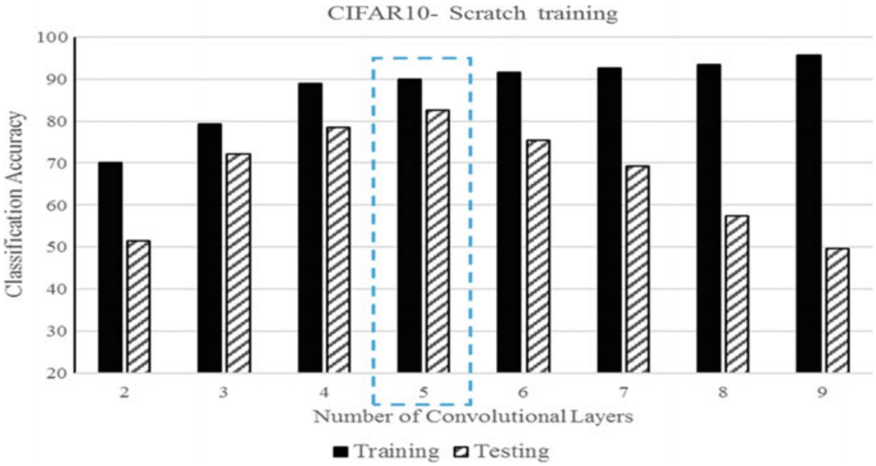


Fig. 6 Scratch training results of CIFAR-10 dataset. Training and testing accuracies are shown in adjacent columns. Best performing architecture is denoted with rectangle

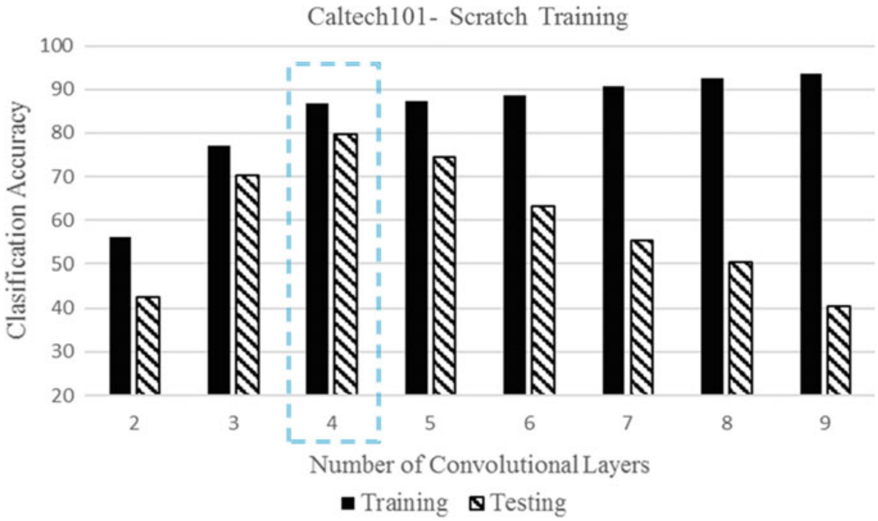


Fig. 7 Scratch training results of Caltech101 dataset. Training and testing accuracies are shown in adjacent columns. Best performing architecture is denoted with rectangle

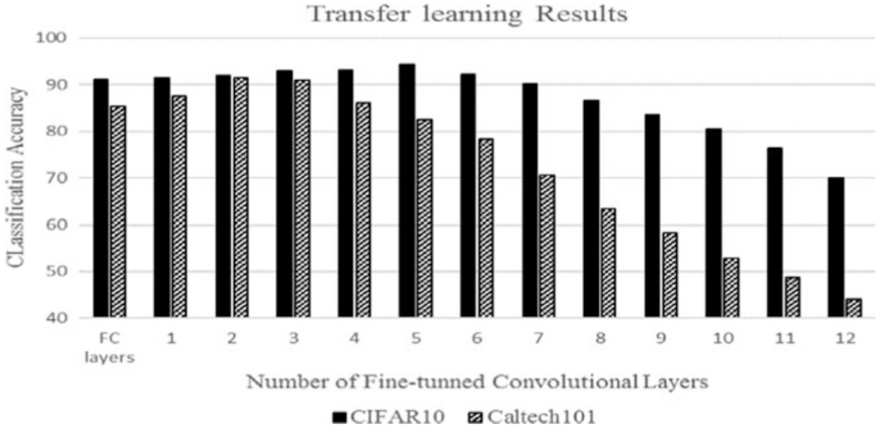


Fig. 8 Transfer learning results of CIFAR10 and Caltech101

performance of transfer learning. The evaluation process starts by re-initializing and training the fully connected layers (this step is denoted as FC layers in Fig. 8) of VGG-16. In the consequent steps, convolutional layers of VGG-16 are fine-tuned (re-initialized and trained) one by one, and corresponding classification accuracies are measured. For example, the fine-tuned convolutional layer three means that the fully-connected layers and last three convolutional layers of VGG-16 are re-initialized and trained.

6 Discussions

Investigating the performance of convolutional neural networks for tiny image classification is important for many practical applications. We conducted this comprehensive study by analysing the performance of scratch (direct) and transfer learning (fine-tuning) based training on two publicly available tiny datasets.

Scratch training results (Figs. 6 and 7) clearly show that while training accuracy is increased with the number of convolutional layers, test accuracy starts falling after some layers. The main reason is that large CNN architectures fail to generalize the tiny data through a huge amount of parameters. Therefore, the model is over-fitted to train data and train accuracy is increased with the number of layers. Because of the over-fitting, the difference between training and testing accuracies is increased in larger architectures. We can also see that CIFAR10 achieves higher classification accuracy with five-layer architecture, while Caltech101 gets higher scratch training accuracy at four layer architecture. Caltech101 is almost seven times smaller than CIFAR10 and therefore, faces the over-fitting issue even at smaller architectures.

Figure 8 clearly shows that transfer leaning based training increases the classification accuracy for both datasets while comparing with scratch training. Since both datasets are much similar with ImageNet (source dataset), they achieve high accuracies with fewer number of fine-tunned layers. Similar to scratch training, accuracies of both datasets are decreased with the number of fine-tunned layers because of the over-fitting issue. Since the Caltech101 dataset fails to train a large number of re-initialized layers, its classification accuracy drops more suddenly than the CIFAR10 for large number of fine-tunned layers.

In transfer learning, layers of source network are frozen, and re-initialized layers are trained with target data. At fine-tuning, Caltect101 achieves high accuracy with an architecture, where the last three convolutional layers are re-initialized. CIFAR10 achieves high accuracy while last five convolutional layers are re-initialized. We investigate these high-performing architectures further. Instead of freezing source layers, we train the whole network (source layer + re-initialized layers) with target data. We have used very low learning rate for source layers (0.0001) and low learning rate (0.001) for re-initialized layers. The results are compared in Table 2.

Table 2 High performing results for different training approaches (HIGH accuracies are shown in bold)

Method	Caltech101	CIFAR10
Scratch Training	79.6	82.5
Fine-tuning (re-initialized layers only)	**91.4**	94.4
Fine-tuning (whole network)	87.8	**95.52**

As shown in Table 2, fine-tuning the whole network increases the accuracy of CIFAR10 to 95.52, while decreasing the accuracy of Catech101 as 87.8. Since Caltech101 is a much smaller dataset for training the whole network (16 layers), fine-tuning the whole network decreases the accuracy considerably. On the other hand, training source network with very smaller learning rate increases the classification accuracy of CIFAR10.

7 Conclusion

In this paper, we conducted a comprehensive study to analyse the classification performance of convolutional neural networks for tiny data. This study was conducted on two publicly available datasets. This study evaluated the performance of two major approaches: scratch training and transfer learning based training. Classification accuracies of scratch training were measured for different architectures by changing the number of convolutional layers. Performance of transfer learning was measured by verifying the number of fine-tuned layers. Based on this study, we can conclude that fine-tuning is the best way to classify the tiny data. It gives high accuracies while the source and target data are much similar. In addition, this study found that accuracies of tiny datasets could be increased by training the source layers with a very low learning rate. These results will be very useful for many practical applications.

References

1. Krizhevsky A, Sutskever I (2012) Hinton GE: ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
2. Lin Y, Lv F, Zhu S et al (2011) Large-scale image classification: fast feature extraction and SVM training. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 1689–1696. https://doi.org/10.1109/CVPR.2011.5995477
3. Wang J, Yang J, Yu K, Lv F, Huang T, Gong Y (2010) Locality-constrained linear coding for image classification. In: 2010 IEEE computer society conference on computer vision and pattern recognition. IEEE, pp 3360–3367
4. Karpathy A, Fei-Fei L (2015) Deep visual-semantic alignments for generating image Des. In: Proceedings of the IEEE conference on computer vision and pattern Recognition, pp 3128–3137. https://doi.org/10.1109/CVPR.2015.7298932
5. Kokul T, Fookes C, Sridharan S, et al (2017) Gate connected convolutional neural network for object tracking. In: IEEE international conference on image processing (ICIP). IEEE, pp 2602–2606 (2017)
6. Kokul T, Ramanan A, Pinidiyaarachchi UAJ (2016) Online multi-person tracking-by-detection method using ACF and particle filter. In: IEEE 7th international conference on intelligent computing and information systems ICICIS, pp 529–536. https://doi.org/10.1109/IntelCIS.2015.7397272

7. Stiller C, Wojek C, Lauer M et al (2013) 3D traffic scene understanding from movable platforms. IEEE Trans Pattern Anal Mach Intell 36:1012–1025. https://doi.org/10.1109/tpami.2013.185
8. Rasheed N, Khan SA (2014) Khalid: a tracking and abnormal behavior detection in video surveillance using optical flow and neural networks. In: Proceeding IEEE 28th international conference on advanced information networking work. IEEE WAINA, pp 61–66. https://doi.org/10.1109/WAINA.2014.18
9. Litjens G, Kooi T, Bejnordi BE et al (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60–88. https://doi.org/10.1016/j.media.2017.07.005
10. Parkhi OM, Vedaldi A, Zisserman A (2015) Deep face recognition. In: Proceedings Br Mach Vis Conference, pp 41.1–41.12. https://doi.org/10.5244/C.29.41
11. Weinberger K (2005) Distance metric learning for large margin nearest neighbor classification. J Mach Learn Res 207–244. https://doi.org/10.1142/S021800141100897X
12. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), pp 770–778
13. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv Prepr arXiv:14091556, pp 1–14. https://doi.org/10.1016/j.infsof.2008.09.005
14. Dalal N, Triggs B (2010) Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). IEEE, pp 886–89
15. Lindeberg T (2012) Scale invariant feature transform. Scholarpedia 7:10491. https://doi.org/10.4249/scholarpedia.10491
16. Jia D, Wei D, Socher R, et al (2009) ImageNet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, pp 248–255. https://doi.org/10.1109/CVPRW.2009.5206848
17. Shin H-C, Roth HR, Gao M et al (2016) Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 35:1285–1298. https://doi.org/10.1109/TMI.2016.2528162
18. Tan C, Sun F, Kong T, et al (2018) A survey on deep transfer learning. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 11141 LNCS: 270–279. https://doi.org/10.1007/978-3-030-01424-7_27
19. Fukushima K (1988) Neocognitron: a hierarchical neural network capable of visual pattern recognition. Neural Netw 1:119–130. https://doi.org/10.1016/0893-6080(88)90014-7
20. Szegedy C, Liu W, Jia Y, et al (2015) Going deeper with convolutions. In: The IEEE conference on computer vision and pattern recognition (CVPR), pp 1–9
21. Simonyan K, Vedaldi A, Zisserman A (2013) Deep inside convolutional networks: visualising image classification models and saliency maps, pp 1–8
22. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 8689 LNCS: pp. 818–833. https://doi.org/10.1007/978-3-319-10590-1_53
23. Bengio Y (2009) Learning Deep architectures for AI. Found Trends® Mach Learn 2:1–127. https://doi.org/10.1561/2200000006
24. CS231n: Convolutional neural networks for visual recognition home page, http://cs231n.stanford.edu/. Last accessed 21 May 2019
25. Cireşan D, Meier U, Schmidhuber J (2012) Multi-column deep neural networks for image classification. https://doi.org/10.1109/CVPR.2012.6248110
26. Cireşan DC, Meier U, Masci J et al (2011) Flexible, high performance convolutional neural networks for image classification. IJCAI Int Jt Conf Artif Intell 1237–1242. https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-210
27. Yosinski J, Clune J, Bengio Y, Lipson H (2014) How transferable are features in deep neural networks? pp 3320–3328
28. Caruana R (1997) Multi-task learning. Kluwer Academic Publishers

29. Chen H, Wang Y, Shi Y et al (2018) Deep Transfer learning for person re-identification. In: 2018 IEEE 4th international conference on big data, BigMM (2018). https://doi.org/10.1109/BigMM.2018.8499067
30. Oquab M, Bottou L, Laptev I, Sivic J (2014) Learning and transferring mid-level image representations using convolutional neural networks. In: 2009 IEEE conference on computer vision and pattern recognition, pp 1717–1724. https://doi.org/10.1109/CVPR.2014.222
31. Sharif A, Hossein R, Josephine A et al (2014) CNN features off-the-shelf-an astounding baseline for recognition. Computer vision and pattern recognition workshops (CVPRW), pp 512–519
32. Guyon I, Dror G, Lemaire V et al (2011) Unsupervised and transfer learning challenge. Proc Int Jt Conf Neural Networks 793–800. https://doi.org/10.1109/IJCNN.2011.6033302
33. Ganin Y, Lempitsky V (2014) Unsupervised domain adaptation by Backpropagation arXiv preprint arXiv:1409.7495
34. Fei-Fei Li, Fergus Rob, Perona Pietro (2007) Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. Comput Vis Image Underst 106:59–70. https://doi.org/10.1016/j.cviu.2005.09.012
35. Krizhevsky A, Hinton G (2009) Learning multiple layers of features from tiny images. In: Learning multiple layers of features from tiny images. University of Toronto
36. Hinton GE, Srivastava N, Krizhevsky A et al (2012) Improving neural networks by preventing co-adaptation of feature detectors arXiv:1207.0580v1[cs.NE], pp 1–18
37. Chollet F (2015) “Keras.”

Development of a WiFi Smart Socket and Mobile Application for Energy Consumption Monitoring

U. A. Ungku Amirulddin, N. F. Ab Aziz, M. Z. Baharuddin, F. H. Nordin and Muhammad Nor Sabrie Johari

Abstract In recent years, with the advancement of Internet-of-Things (IOT), there has been research into the development of smart sockets that are able to monitor the energy consumption of appliances. Some smart sockets can only monitor energy consumption, while others include a controller, which is able to make decisions to switch on/off the connected appliance without user interaction. This paper presents a first prototype of a smart socket based on an advanced 32-bit ESP32 Microcontroller Unit (MCU), which features an in-built Wi-Fi 802.11 b/g/n connection. The MCU was programmed and interfaced with devices to enable voltage, current, power and power factor measurement, which are transmitted to a cloud-based server. A mobile application was also developed using the Blynk platform, which enables display of the connected appliance energy consumption data. The application also enables the user to remotely switch on/off the appliance through the MCU. It is hoped that this design will be a starting point towards embedding further improved features on the smart socket, which will enhance efforts to improve energy efficiency of households.

Keywords Smart socket · Energy efficiency · WiFi · IOT

1 Introduction

Energy efficiency is becoming increasingly popular in this decade due to increased efforts globally to conserve energy usage. The latest report by International Energy Agency (IEA) states that the global energy demand in the year 2017 had increased by 1.9% as compared to the previous year and this was recorded as the fastest annual increase since the year 2010 [1]. This is due to strong economic growth that

U. A. Ungku Amirulddin (✉) · N. F. Ab Aziz · M. Z. Baharuddin ·
F. H. Nordin · M. N. S. Johari
Institute of Power Engineering (IPE), Universiti Tenaga Nasional,
Jalan IKRAM-UNITEN, 43000 Kajang, Selangor, Malaysia
e-mail: anisa@uniten.edu.my

Z. Zakaria and R. Ahmad (eds.), *Advances in Electronics Engineering*,
Lecture Notes in Electrical Engineering 619,
https://doi.org/10.1007/978-981-15-1289-6_10

was outpaced by progress on energy efficiency. However, energy efficiency activities implemented worldwide since the year 2000 have caused an impact by preventing 12% more energy use in 2017 [1].

One of the key sectors for energy efficiency is in buildings and appliances. Policies are in place to ensure that new buildings implement energy efficiency. However, buildings built prior to the implementation of the policies are encouraged to be retrofitted in order to ensure energy efficiency. Some examples of retrofitting which can be done are the use of photovoltaics on rooftops and using energy efficient lighting as well as appliances. Unfortunately, retrofitting efforts can incur large upfront costs to the household owner. Thus, another possible method to improve energy efficiency is through monitoring of electricity consumption within the household. A typical household consists of many appliances that are connected to socket points throughout the house. In recent years, there has been advancement in the development of smart sockets or switches that are able to monitor the energy consumption of appliances.

Shie et al. [2] had developed a smart socket which was capable of measuring voltage, current and power factor of an appliance. The data measured are then transmitted via a short-range wireless Zigbee network protocol to a monitoring server within the household. A user interface sorts the power usage statistics according to date and time. A system to measure energy consumption at the mains socket consisting of several Smart Energy Measuring Devices (SEMD) was also developed by Altmann et al. [3]. Voltage and current measurements are taken from the appliance connected to the SEMDs. The measurement samples are averaged and communicated to a gateway every minute and sent to a server using 868 MHz radio transmission with a self-developed bi-directional low-power protocol. Data stored can then be accessed via web interface. However, the smart sockets developed by Shie et al. [2] and Altmann et al. [3] are purely for energy consumption monitoring. The appliances are not controlled to switch on or off to improve energy usage efficiency.

A home energy saving network structure consisting of a Digital Power Meter Unit (DPMU), a Metering Interface Unit (MIU), a Data Center Unit (DCU) and a Smart Plug Unit (SPU) was developed by Chen and Lin [4]. The SPU is able to monitor load current of the appliance connected in real time. The data is then transferred to the MIU via a Zigbee network. The MIU will then decide to either switch off the appliance or keep it switched on based on the power consumed. In the instance that the appliance is switched off, the MIU will send a message to the user's mobile phone via a wireless module. Pawar and Vittal [5], on the other hand, designed a smart socket to be integrated within a Home Energy Management System (HEMS) with a master controller. When first connecting an appliance to the smart socket, a request to the master is sent via a Zigbee network. Upon receiving the request from the smart socket, the master runs a power negotiation algorithm to figure out a power budget, and then approves or rejects the request. On approval, the appliance gets turned on and the smart socket starts measuring the energy consumed by the appliance until the appliance is turned off. Apart from measuring energy consumed, the socket developed by Pawar and Vittal [5] also measures different electrical parameters, such as power factor, average power, apparent power, and Total Harmonic Distortion (THD). A similar smart socket was also developed by Al-Hassan et al. [6], which measures RMS voltage and current, power, and power factor of a connected appliance and transmits the data to a master controller through a Zigbee network. The master controller will then decide on whether to keep the appliance on or switch it off. Thus, the smart sockets developed by Chen and Lin [4], Pawar and Vittal [5] and Al-Hassan et al. [6] employ a master controller to make decision on whether to remotely switch off the appliance connected to the socket without the interaction of the user.

With recent advances in Internet Of Things (IOT), Tsai et al. [7] combined the smart socket and IOT in their Residence Energy Control System (RECoS). The RECoS has an automatic control mode, which turns on/off the power of the smart socket for particular time periods or when the total energy consumption of the smart socket exceeds a user predefined limit. The RECoS also has a user control mode, whereby the user can remotely send command from the user's smartphone or PC to the smart socket on/off switch based on the electricity data of the connected appliance, i.e. voltage, current, power, accumulated energy, phase, and frequency. The smart socket designed by Musleh et al. [8] can also be remotely switched on or off using a Windows-based phone application platform based on the recorded energy consumption of the connected appliance. The voltage and current measurements from the smart socket [8] are relayed via a Zigbee network to a Raspberry Pi master unit, which controls the slave sockets. The master will also send the data measured to a web server.

In this paper, we propose a smart socket, which is able to monitor voltage, current, power and power factor of a connected appliance and transmit the measured data to a cloud-based server via WiFi. The smart socket is accompanied by a mobile application that is able to display the energy consumption of the appliance and enables the user to remotely switch off the connected appliance based on the energy consumed. The following section will describe the proposed smart socket design. Then, the mobile application will be explained before the paper is concluded.

2 Proposed Smart Socket Design

Figure 1 shows the proposed design of the smart socket. The smart socket is to be placed in between the home mains socket and the AC appliance which is to be monitored. A measurement unit is required within the smart socket to measure the voltage, current, active power, power factor, frequency, and active energy of the connected AC appliance. The voltage measurement is obtained through the Live (L) and Neutral (N) inline feed from the mains socket that is connected with the appliance. The current flow throughout the appliance is measured using a current transformer, which is placed non-invasively on the Neutral line of the appliance. The measured data is then relayed to the main controller through a RS485

Modbus RTU protocol to UART TTL Serial protocol converter. The main controller of the smart socket is a 32-bit ESP32 Microcontroller Unit (MCU). The MCU is powered with 9 V DC voltage that is obtained from a DC-DC buck converter, which steps down the 12 V DC power supply unit from the 240 V AC mains socket. Several LED's and a buzzer are used to provide a visible and audible feedback for the user.

The MCU will perform mathematical calculations from the measured data for the appliance energy consumption monitoring. The energy consumption data will then be encoded by the MCU and transmitted to a cloud-based server, with a specific API key for credential purposes, through the built-in Wi-Fi 802.11 b/g/n internet connection. The user will then be able to view the energy consumption data wirelessly through a mobile application and decide on whether to keep the appliance on or switch it off. If the user decides to switch off the appliance remotely, through the mobile application, the signal will be transmitted to the MCU via internet connection and decoded by the MCU as a control signal. The control signal will be conveyed from the MCU to the solid-state relay which will switch off the appliance. The hardware circuit constructed for the smart socket is shown in Fig. 2.

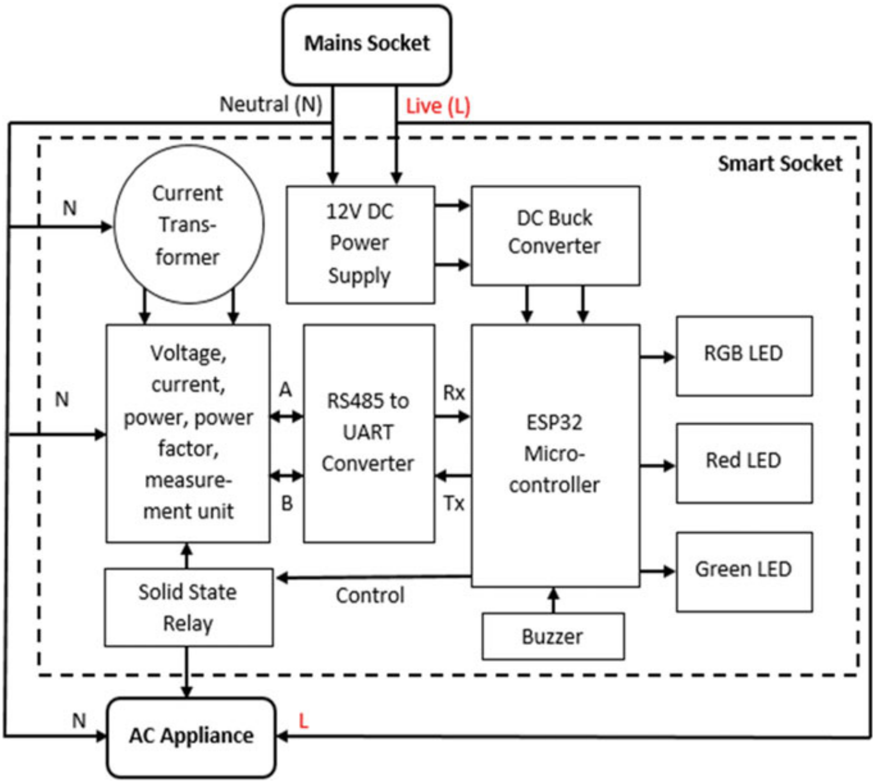


Fig. 1 Smart socket system diagram

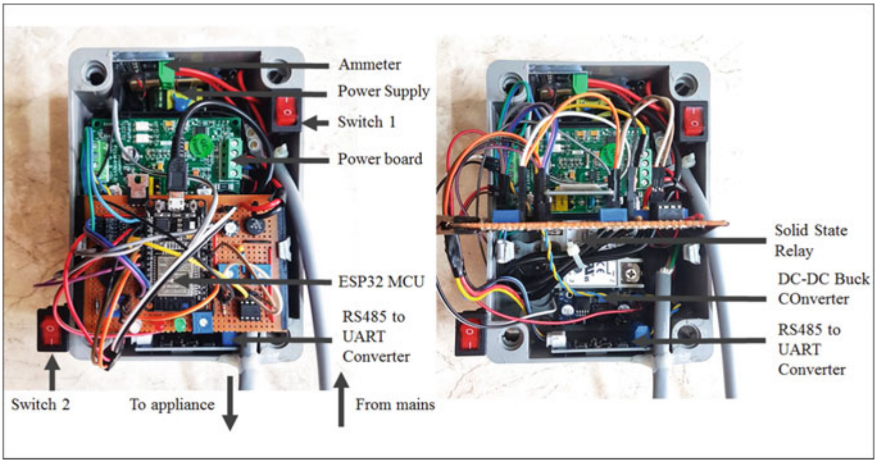


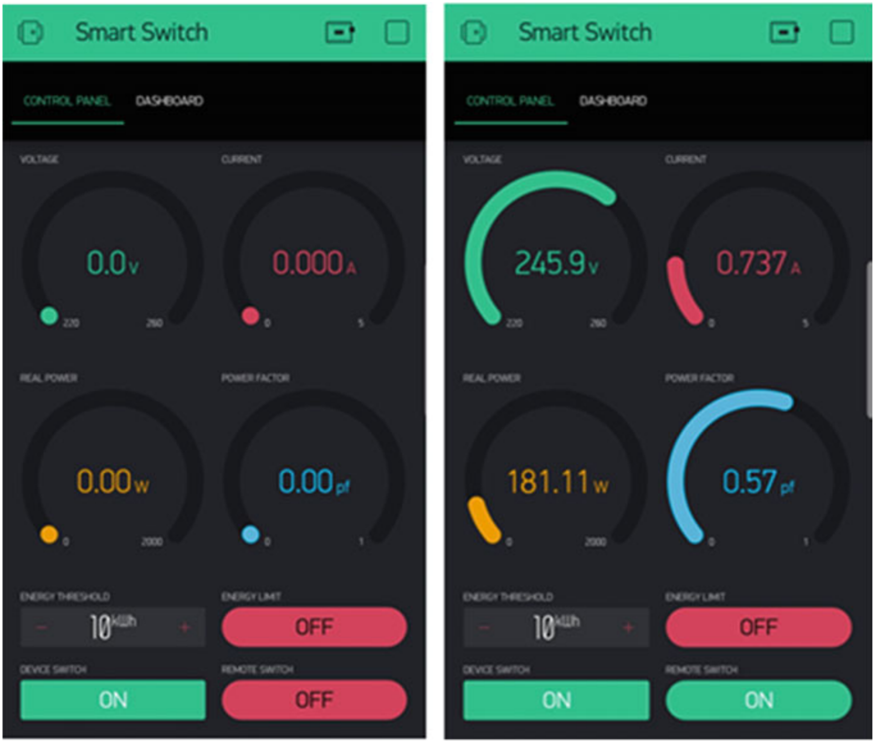
Fig. 2 Smart socket circuit

3 Mobile Application Development

The smart socket is to be used with a mobile application developed using the Blynk platform. The interface of the mobile application is divided into two sections: 'Control Panel' and 'Dashboard'. The 'Control Panel' section of the mobile application enables the user to monitor the real-time measurement of the appliance connected to the smart socket (i.e. voltage, current, power, and power factor). The 'Remote Switch' on the mobile application can be used by the user to remotely switch on or off the appliance from the mains power supply. However, this can only be done if and only if the physical Device Switch at the smart socket is turned on. Toggling the 'Remote Switch' button on the mobile application while the physical Device Switch is off would have no effect on the connected appliance, thus, causing no power to be supplied to the appliance from the mains socket. Figure 3 shows the information displayed in the 'Control Panel' section of the mobile application.

The user is also able to limit the energy consumption of the connected appliance through the mobile application by toggling on the 'Energy Limit' button. The value of the energy consumption limit can be set by the user within the range of 0–9999 kWh on the 'Energy Threshold' input box. The supply to the appliance will be automatically disconnected once the energy consumption exceeds the preset limit. The total energy consumption of the appliance can be reset to zero by toggling off the Device Switch on the mobile application.

The 'Dashboard' section on the interface of the mobile application will allow the user to select which smart socket to be monitored and controlled by the application, if there are more than one smart socket being used by the user in the household. The user can also review historical data in graphical form based on hourly, daily, monthly, or yearly timeframes. Figure 4 shows two examples of the historical data



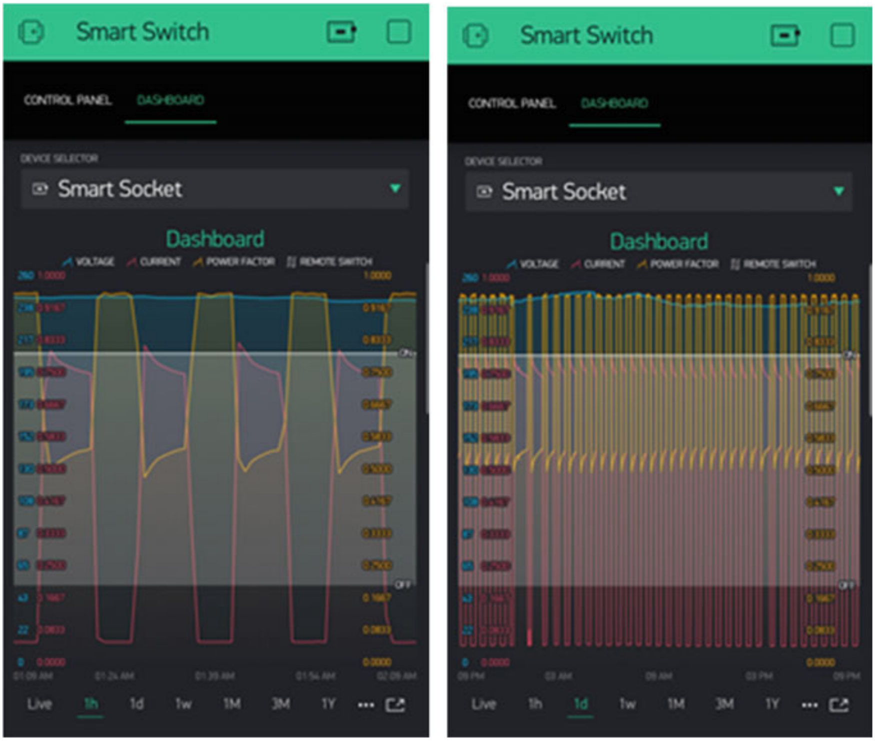
(a) Remote Switch turned OFF (b) Remote Switch turned ON

Fig. 3 ‘Control Panel’ section of the mobile application

given by the ‘Dashboard’ section of the mobile application on hourly and daily timeframes. In this figure, electricity consumption of a home refrigerator unit was measured. The compressor cycles can clearly be seen, which in turn affects the power factor.

4 Conclusion

Energy efficiency is increasingly becoming an important requirement in order to curb global warming. Within a typical household, a user may employ the use of smart sockets to monitor energy consumption and improve energy efficiency of the household. This paper presented the design of a smart socket, which is accompanied by a mobile application to monitor the energy consumption of an appliance connected to the smart socket. The smart socket is placed in between the mains socket and the AC appliance which is to be monitored. The measured voltage, current, power, and power factor are transmitted via WiFi to a cloud-based server



(a) 1-Hour timeframe

(b) 1-Day timeframe

Fig. 4 The ‘Dashboard’ historical data displayed at different timeframes

from the 32-bit ESP32 Microcontroller Unit. The data can be monitored wirelessly through the mobile application, which was developed using the Blynk platform. Based on the displayed energy consumption, the user can remotely switch off the appliance through the mobile application. The designed smart socket and mobile application are the first prototypes and further enhancements can be made to enable more features to be added, which can enable the smart socket to provide more intelligent decisions that can aid the user to improve energy efficiency within the household.

Acknowledgements This project is part of the Smart University project funded by Universiti Tenaga Nasional (UNITEN). The authors would like to acknowledge UNITEN for the funding provided for the project. Acknowledgement is also extended to Blynk for the development of the mobile application.

References

1. Market Report Series (2018) Energy efficiency 2018 analysis and outlooks to 2040, international energy agency
2. Shie M, Lin P, Su T, Chen P, Hutahaean A (2014) Intelligent energy monitoring system based on zigbee-equipped smart sockets. In: Proceedings of international conference on intelligent green building and smart grid (IGBSG). Taiwan, pp 1–5
3. Altmann M, Schlegl P, Volbert K (2015) A low-power wireless system for energy consumption analysis at mains sockets. In: Proceedings of 12th international workshop on intelligent solutions in embedded systems (WISES). Italy, pp 79–84
4. Chen M, Lin C (2015) Design and implementation of a smart home energy saving system by multi-microprocessor. In: Proceedings of international conference on consumer electronics-Taiwan. Taiwan, pp 410–411
5. Pawar P, Vittal KP (2017) Design of smart socket for power optimization in home energy management system. In: Proceedings of 2nd IEEE international conference on recent trends in electronics information & communication technology (RTEICT). India, pp 1739–1744
6. Al-Hassan E, Shareef H, Islam MM, Wahyudie A, Abdrabou AA (2018) Improved smart power socket for monitoring and controlling electrical home appliances. IEEE Access 6:49292–49305
7. Tsai K, Leu F, You I (2016) Residence energy control system based on wireless smart socket and IoT. IEEE Access 4:2885–2894
8. Musleh AS, Debouza MI, Farook M (2017) Design and implementation of smart plug: an internet of things (IoT) approach. In: Proceedings of international conference on electrical and computing technologies and applications (ICECTA). United Arab Emirates, pp 1–4 (2017)

Forecasting International Tourist Arrivals from Major Countries to Thailand

Ontheera Hwandee and Naragain Phumchusri

Abstract Tourism industry is one of the most important industries for Thai economy. This paper proposes and compares forecasting models for international tourism arrivals to Thailand. Since country-specific forecasting models can reflect the uniqueness of each country of origin, major countries for Thai tourism, namely China, Malaysia, Korea, Japan, and Russia are explored. The data used in this research is the number of international tourist arrivals from those countries recorded monthly from Jan 2013 to Sep 2018. The performance of the Seasonal Autoregressive Integrated Moving Average model (SARIMA) and the multiple regression model are evaluated in terms of Mean Absolute Percentage Error (MAPE). Several important economic factors such as income, price, exchange rate, and qualitative factors, represented by dummy variables of seasonal effect are explored to understand their effects on international tourism demand. The results show that the SARIMA is preferred to forecast international tourism arrivals from Malaysia, while multiple regression provides lowest errors for other interested countries.

Keywords Forecasting · Tourism industry · Tourist arrivals · SARIMA · Multiple regression

1 Introduction

The tourism industry is one of the world's largest industries with a global economic contribution (direct, indirect, and induced) of over 7.6 trillion U.S. dollars in 2016. The direct economic impact of the industry, including accommodation, transportation, entertainment, and attractions, was approximately 2.3 trillion U.S. dollars

O. Hwandee (✉) · N. Phumchusri
Department of Industrial Engineering, Chulalongkorn University, Bangkok, Thailand
e-mail: onhwandee@gmail.com

N. Phumchusri
e-mail: naragain.p@chula.ac.th

Z. Zakaria and R. Ahmad (eds.), *Advances in Electronics Engineering*,
Lecture Notes in Electrical Engineering 619,
https://doi.org/10.1007/978-981-15-1289-6_11

in that year. A number of countries, such as France and the United States, are consistently popular tourism destinations, but other less well-known countries are quickly emerging in order to reap the economic benefits of the industry.

Thailand is one of the countries having many natural and cultural tourism products and beautiful beach destinations. All these elements profoundly attract tourists from all over the world. Moreover, Thailand is consistently ranked in the top ten for its beaches, entertainment and dining, value of products, recreational facilities, and shopping [1]. Therefore, tourism is a very important industry to Thailand's economy. It contributes significantly to Thailand's Gross Domestic Product (GDP), affecting employment, investment, and foreign exchange earnings.

The number of international tourists in Thailand increased from 10.87 million in 2002 to 35.38 million in 2017 shown in Fig. 1. Direct contribution of travel and tourism to GDP was worth 455.22 billion US dollars in 2017 [2]. In addition, the tourism market in the form of Thai culture to attract foreign tourists to travel in Thailand increased in 2018. It is important for the economy; policy makers should pay attention to this demand. Determining the level of tourist demand helps planners reduce the risk of decisions regarding the future. It also provides information for tourism suppliers to prepare suitable products for each group of tourists.

To comprehend at the best the tourism demands, several studies have been published in order to understand the demand and its components. It is vitally important to forecast tourism demand in the region and understand the factors affecting demand. Thus, country-specific forecasting models and strategies should be formulated to reflect the uniqueness of each country of origin. Furthermore, forecasting techniques should include more qualitative factors to better asses their impacts on tourism demand. Thus, the objective of this research is to develop and compare forecasting models to forecast international tourism arrivals from major countries to Thailand.

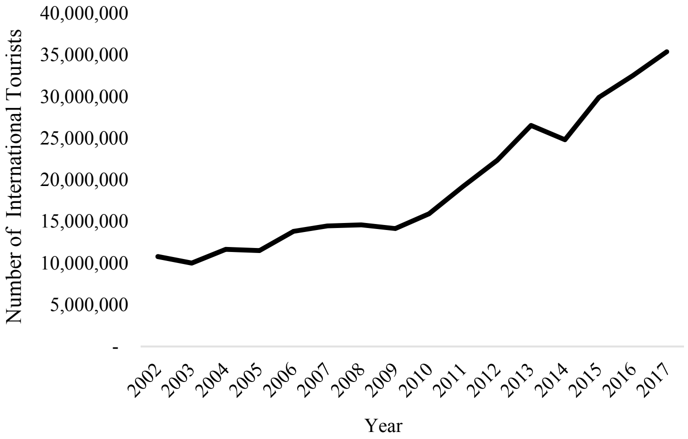


Fig. 1 The international number of tourists visited Thailand from 2002–2017

2 Literature Review

2.1 Summary of Explanatory Methods for Tourism Demand

Tourism demand can be measured in terms of the number of tourist arrival in most studies [3], tourist arrivals have been used as a dependent variable. Variation in tourism inflows is induced by many factors. Many studies focused predominantly on economic factors such as income, relative prices, and exchange rates.

Income Income in the origin country is the most frequently used explanatory variable in the published tourism studies. The Most common proxies for income that are used in tourism research are national income in the form of Gross Domestic Product (GDP) and Gross National Product (GNP) [4]. However, the GDP was the preferred measure of income and positively related with the number of tourists [1, 5, 6].

Price Consumer price are costs of goods and services that tourists are likely to pay while at the destination (such as accommodation, local transportation, food, and entertainment). Most previous studies used relative price in terms of relative Consumer Price Index (CPI) of that country [7, 8, 9]. The calculation of tourism price is based on the Consumer Price Index (CPI) of the destination divided by the CPI of the country of origin that is shown in this equation.

$$\text{Relative Prices}\frac{\text{CPI(Destination)}}{\text{CPI(Origin)}} \tag{1}$$

Exchange Rate Many studies specifically examine impact of exchange rates on international tourism demand. The exchange rates have some influences on destination choice and travel purchases. The exchange rate factor became an important factor in terms of economy [10, 11, 12]. However, the exchange rate can be both positive and negative and this relies on relative value of the based country [13].

Other Factors Dummy variables may be used to capture seasonal variations in tourism demand. Seasonal patterns in tourist flows and expenditures are well-known characteristics of international tourism demand. Specific time of the year, like a season or a period of school holidays, can have a significant effect on tourism demand. Seasonality has been dealt with by many authors but has been avoided by some due to modelling tourism demand based on annual data. Typically, if using monthly data, twelve seasonal dummy variables are included in the model and similarly four seasonal dummy variables are incorporated regarding the quarterly data [13].

2.2 Summary of the Time Series Model for Tourism Demand

Time series models have been widely used for tourism demand forecasting with the integrated autoregressive moving-average models (ARIMAs) proposed by [14]. This model could be used with the latter gaining an increasing popularity over the last few years. GARCH model has been applied in terms of economic model [15]. Moreover, other time series methods have also appeared frequently in many studies such as Naïve, exponential smoothing models, and simple autoregressive.

2.3 Review of Forecasting Demand for Thai Tourism

The existing literature on forecasting tourism demand in Thailand is usually in terms of the different techniques employed. The model selection is often based on the out-of-sample forecasting performance. The forecasting techniques in terms of the different countries covered [16], seven major origin countries—Australia, Japan, Korea, Singapore, Malaysia, the UK, and the USA– are examined with classical regression analysis. Economic theory is used to recommend which variables should be income, own price, cross price, and trade volume. Furthermore, qualitative factors are included [1, 17]. In recent years, there are also several forecasting models for the international tourist arrivals to Thailand in terms of technical comparison including those by Box-Jenkins, regression analysis, and Brown's double exponential smoothing [18], comparing with the Box–Jenkins and Winter's methods [19].

From existing literature, it can be observed that the forecasting and modelling of international tourist arrivals to Thailand from major countries have not been considered in term of comparison between time series model and explanatory model. Moreover, seasonal effect has not yet been included in the previous models. On the other hand, this research focuses on capturing seasonality factors by comparing the multiple regression (focusing on economic factors and dummy variable of seasonal effect) with the SARIMA model. This can help to explain the change in tourism demand in the short-run, and monthly arrivals can be predicted.

3 Research Methodology

The objective of this paper is to propose forecasting models that can accurately estimate the international tourism demand for Thailand. To handle the increasing variety and complexity of managerial forecasting problems, many forecasting techniques have been developed in recent years. Each has its special use, and care must be taken to select the correct technique for a particular application.

The selection of a method depends on many factors: the context of the forecast, the relevance and availability of historical data, the degree of accuracy desirable, the time period to be forecast, the cost/benefit (or value) of the forecast to the company, and the time available for making the analysis.

3.1 The Data

This study also aims to analyse historical data on the international tourist arrivals in Thailand in order to identify any pattern or trends for providing insights for future.

To investigate the trend and seasonality of international tourist arrivals to Thailand with monthly data, the time series plot of international tourist arrivals of each country would provide insights. In this and following sections, monthly data between Jan 2013 and Sep 2018 are of interest.

The data are divided into 2 intervals for training and testing the model. The monthly data in 2013 to 2017 were selected to construct the model and the rest are used for testing.

The factors that impact on the number of tourist arrivals to Thailand are studied in economic term. The frequently used variables in the literatures are income that measure in term of GDP(USD), price that measure in term of monthly relative CPI and monthly exchange rate. The seasonal dummy variables such as yearly effect and monthly effect are included in the model. All the historical data of economic factors obtained from economic trader website.

Time series plot from Fig. 2 shows an example of time series data of the number of Chinese tourist arrivals to Thailand from Jan 2013 to Sep 2018, the pattern of time series clearly shows there are trends and seasonality components.

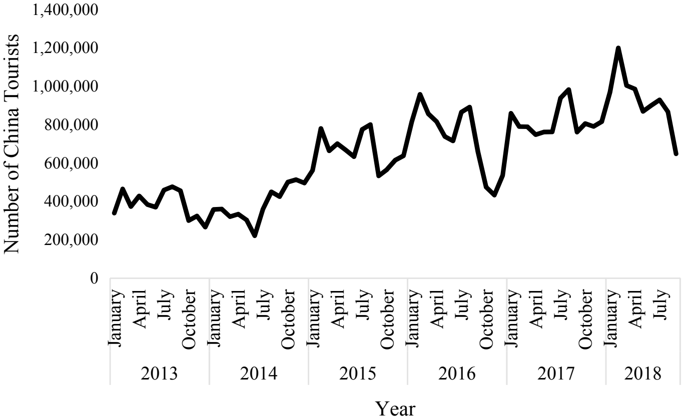


Fig. 2 The number of Chinese tourist arrivals to Thailand in Jan 2013 to Sep 2018

3.2 Forecasting Models

Two quantitative methods are presented in this study as follows; Causal forecasting method and Time series method.

Causal Forecasting Models Causal forecasting models usually consider several independent variables that are related to the dependent variable being predicted. The most common quantitative causal forecasting method is linear regression analysis. In this study, economic explanatory factors are used as independent variables. The general form of a multiple regression model is:

$$Y_t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon \quad (2)$$

where

X_1 is income in terms of GDP per capita (billion USD)
X_2 is relative price
X_3 is exchange rate
X_4 is dummy variable of seasonal effects
β_0 is constant variable
β_1 is coefficient of the first control variable, X_1
β^2 is coefficient of the second control variable, X_2
β_3 is coefficient of the third control variable, X_3
β_4 is coefficient of the forth control variable, X_4
ε is error term.

The dependent variable is the number of tourists visited Thailand from each country.

The demand for tourism is a function of its determinants as follows:

$$Q_{ij} = f(G_i, C_i, E_{ij}, D_j \varepsilon_{ij}),$$

where:

Q_{ij} quantity of tourism demanded in destination i by tourists from country j,
E_{ij} exchange rate from origin country j to destination country i.
C_{ij} price of tourism for destination i divided by original country j.
G_{jt} GDP from origin country j in period t.
D_{it} Seasonal effect, set of k-1 monthly seasonal dummy variables where k is an indicator of the month ($k = 1 \sim 12$) to determine the seasonal pattern of tourist arrivals from the source i in period t. For this study, I exclude the year of 2014 and month of September. Consequently, all other monthly (seasonal) effects will reflect the difference from the September effect.

Time Series Models ARIMA models are commonly used in tourism forecasting for time series. This is known as the ARIMA (p, d, q) model, where d denotes the number of times a time series has to be differenced to make it stationary. For seasonal time series, literature suggests seasonal autoregressive integrated moving average models, also called SARIMA (p, d, q) (P, D, Q) s models. In this paper, we focus on SARIMA model, which can be expressed as:

$$\Phi(L^s)\phi(L)\Delta^d\Delta_s^d y_t = \theta_0\Theta(L^s)\theta(L)\varepsilon_t \tag{3}$$

where s is the seasonal length, for example, s = 12 for monthly and s = 4 for quarterly data, L is the lag operator and Δ_t is assumed to be a Gaussian white-noise process with mean zero and variance σ^2. The difference operator is Δd, where d specifies the order of differencing and the seasonal difference operator is ΔD_s where D is the order of seasonal differencing. The difference operators are applied to transform the observed non-stationary time series y_t to the stationary.

Fitting Box–Jenkins Models for A Seasonal Model. A seasonal model is identified using the following steps

Step 1 Examine the time-series plots for seasonality and trend (i.e., check for stationarity).

Step 2 If data exhibit trend and seasonality effects, turn the data into a stationary series, use both seasonal and nonseasonal differencing and apply.

Step 3 Examine the ACF and PACF of the transformed data to identify the AR and MA terms. The best model was picked with the p-value of parameter less than 0.05.

Step 4 When the model is selected, its parameters can be estimated using least-squares method.

Step 5 Perform tests on the residuals in order to determine whether the model is adequate for the data. It is necessary to check the assumptions of normality and autocorrelations (using the Ljung–Box test). Finally, the model's in-sample and out-of-sample forecasting accuracy are tested.

4 Results and Discussion

4.1 Multiple Regression Model

The factors influencing international tourist flows to Thailand from major countries were analysed using the ordinary least squares with multiple regression technique. The results for all 5 countries are shown in Table 1. Only those variables with a significant p-value of 0.05 or better are included in the final equations reported. The table is divided into four parts. The top part shows constant term, and macroeconomic variables affecting number of tourist arrivals. The second and third part

Table 1 Estimated models for top five countries arrivals to Thailand 2013–2017

Equation	China	Malaysia	Korea	Japan	Russia
Constant	8,771,675	997,045	−991,763	−1,828,972	−1633
GDP			38.57	60.41	
Relative CP I	−8,284,103	−858,913		−368,169	
Exchange rate					95,394
Yearly effect					
2013		45,320	88,505	−116,250	
2015	88,822		48,115	217,353	−34,085
2016		35,594	39,129	−42,838	−13,588
Seasonal effect					
Jan		−24,764	51,763		132,342
Feb	103,425		23,972		98,663
Mar	98,824				94,858
Apr	116,713		−12,007	−20,462	44,244
May	124,764		−11,632	−25,478	
Jim	130,985			−25,611	
Jul	255,326		18,308	−15,277	
Aug	218,201		32,866	24,471	
Oct				−21,777	51,459
Nor			14,106		111,078
Dec		94,756	34,000	5235	135,263
R-square	89.01	73.82	90.84	92.5	92.16
R-square adj	86.41	70.85	88.5	90.58	90.56

shows yearly seasonal effects and monthly effect of tourists from each source country that 2014, September is used as the base. And the fourth part show how much the estimated equation can be explained the relation.

From Table 1, it is found that only Russia has a significant exchange rate term. It results from the weak exchange rate between the Russian Ruble and other currencies. Therefore, Russian tourists face with higher costs in terms of the Ruble. Seasonal factors largely reflect the holiday pattern for each source country. For most countries, December is the peak travel season for tourists except China. The majority of the individual models had a relatively high R^2 with significant coefficients. It means that the probability of multicollinearity existing was minimal.

4.2 Seasonal ARIMA Model

After approaching the examination of time series plot, the data is not stationary. Thus, data are transformed using differencing to stationary data. Differencing for non-seasonal and seasonal terms were performed in Table 2.

Table 2 Summarized results of SARIMA model of each country

Country	China	Malaysia	Korea	Japan	Russia
Modelling	(0,1,0) $(1,0,0)_{12}$	(2,0,0) $(1,1,0)_{12}$	(0,1,0) $(0,1,0)_{12}$	(0,1,0) $(0,1,1)_{12}$	(0,1,2) $(0,1,1)_{12}$
MAPE training	11.56	7.9	4.5	5.8	8.8
MAPE testing	14.4	9.7	5.1	9.5	31.0

Based on Table 2, the models were compared using MAPE. The MAPE values, which are less than 10%, indicate that the forecast value can also be considered as accurate. After applying the forecasting model to those countries, the error for arrivals from China and Russia models are still high, while the performance for other countries are satisfying (with MAPE < 10%).

However, in order to obtain the best model to forecast tourism data for each country, the model performance was compared between SARIMA and multiple regression using Mean Absolute Percentage Error (MAPE) as shown in Table 3.

From Table 3, the parameters obtained from training data in Jan 2013 to Dec 2017 are used to forecast the number of tourists in Jan to Sep 2018. The results show the multiple regression model performs better than SARIMA in forecasting the number of tourists from China, Japan, Korea and Russia. The reason for this is probably due to the fact that estimated model explicitly addresses causal relationship between the number of tourists from those countries and relative CPI, GDP, and seasonal dummy variables. Meanwhile, the SARIMA model is more adequate in forecasting the number of tourists from Malaysia.

However, MAPE tested value of China and Russia are more than 10%. It may be because the current time series pattern has changed. In order to achieve strategic planning for those countries, other insides should be obtained from other sources. Further advanced models can be explored if they can help improve accuracy for those counties.

Table 3 Comparison of residual errors between SARIMA and multiple regression model

Country	Modelling	MAPE test
China	SARIMA	14.4
	MR	13.2
Malaysia	SARIMA	9.7
	MR	9.8
Japan	SARIMA	5.1
	MR	2.61
Korea	SARIMA	9.5
	MR	8.9
Russia	SARIMA	31
	MR	24.13

Since, all economic factors are related with number of tourist arrivals in Thailand, it is important to analyze how each factor is important. Moreover, dummy variable, e.g. seasonal effect, may not be able to explain the pattern completely. Therefore, the future work will include the event that impact on number of tourists.

5 Conclusions

This research proposes and compares forecasting models for international tourism arrivals to Thailand focusing on five major origin countries. The historical data were analysed using the general-to-specific modelling methodology and the key determinants were identified. Forecast values were compared based on the mean absolute percentage error. The result showed that the SARIMA is preferred to forecast international tourism arrivals from Malaysia, while multiple regression provides lowest errors for other interested countries. The numbers of tourist arrivals from those countries are sensitive to income, price variable, and seasonal dummy variables.

It is interesting to extend our result to other main countries such as India, USA, and England to explore how the models perform. Another extension on improving the forecasting methods such as applying Artificial Neuron Network (ANN) or the hybrid ARIMA-ANN model can be interesting. It is also possible to extend this work to forecast international arrivals to each main province in Thailand to better understand the patterns of data in those areas.

References

1. Sookmark S (2011) An analysis of international tourism demand in Thailand, School of Development Economics (2011)
2. International Monetary Fund (2018) Gross domestic product of Thailand
3. Lim C (1997) Review of international tourism demand models. Ann Tour Res 24:835–849
4. Lim C, Mcaleer M (2002) Time series forecasts of international travel demand for Australia. Tourism Management, 389–396
5. Untong A, Ramos V, Kaosaard M, Reymaquieira J (2015) Tourism demand analysis of Chinese arrivals in Thailand. Tourism Economics 6:1221–1234
6. Song H, Li G, Witt S (2010) Tourism demand modelling and forecasting: how should demand be measured. Tour Econ 1:63–81
7. Salman AK (2013) Estimating tourism demand through cointegration analysis. Curr Issues Tour 6:323–338
8. Lim C (2004) The major determinants of korean outbound travel to Australia. Math Comput Simul 64:477–485
9. Dritsakis N (2004) Cointegration analysis of german and british tourism demand for Greece. Tour Manag 25:111–119
10. Kah J, Lee SH (2013) The value of Japanese Yen and Japanese tourism in Korea. Int J Digit Content Technol Its Appl (JDCTA) 7:302–306

11. Agiomirgianakis G, Serenis D, Tsounis N (2015) Effects of exchange rate volatility on tourist flows into Iceland. Procedia Econ Financ 24:25–34
12. Hanafiah M, Harun M (2010) Tourism demand in Malaysia: a cross-sectional pool time-series analysis. Int J Trade 1:80–83
13. Lean HH, Chong SH, Hooy CW (2014) Tourism and economic growth: comparing Malaysia and Singapore. Int J Econ Manag 8:139–157
14. Box G, Jenkins GM (1970) Time series analysis, forecasting and control
15. Chan F, Lim C, McAleer M (2005) Modelling multivariate international tourism demand and volatility. Tour Manag 26:459–471
16. Song H, Stephen F, Ang G (2003) Modelling and forecasting the demand for Thai tourism. Tour Econ 4:363–387
17. Hao J, Var T, Chon J (2003) A forecasting model of tourist arrivals from major markets to Thailand. Tour Anal 8:33–45
18. Boonaom N (2018) Forecasting the number of chinese tourists in Thailand. Thammasat J 26
19. Luckana S, Sunee T, Yupin, K, Boonying S (2014) A forecasting methods for the number of international tourists in Thailand: box-jenkins method and winter's method. University of the Thai Chamber of Commerce

Depth Estimation Based on Stereo Image Using Passive Sensor

Rostam Affendi Hamzah, M. G. Y. Wei, N. S. N. Anwar, S. F. Abd Gani, A. F. Kadmin and K. A. A. Aziz

Abstract This article presents an algorithm for depth estimation using a pair of passive sensors which involves two cameras. These cameras did not produce any energy to collect the depth information. However, the depth information obtained from a camera can be produced by a matching process between two images at the same viewpoints. These images are captured from two cameras, which are also known as stereo cameras. The matching process consists of several stages, which will produce depth map. The most challenging problem for the matching process is to get an accurate corresponding point between two images. Hence, this article proposes an algorithm for stereo matching using Weighted Sum of Absolute Differences (WSAD), Median Filter (MF), and Bilateral Filter (BF) to surge up the accuracy. The WSAD will be implemented at the first stage to get the preliminary corresponding result, then the BF works as an edge-preserving filter to remove the noise from the first stage. The MF is used at the last stage to improve final depth map. A standard benchmarking dataset from the Middlebury has been used for the experimental analysis and validation. The proposed work in this article achieves good accuracy. The comparison is also conducted with some established methods where the proposed framework performs much better.

Keywords Weighted sum of absolute differences · Median filter · Stereo matching algorithm · Stereo vision · Bilateral filter

R. A. Hamzah (✉) · S. F.Abd Gani · A. F. Kadmin · K. A. A. Aziz
Fakulti Teknologi Kejuruteraan Elektrik dan Elektronik, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian, Tunggal, Melaka, Malaysia
e-mail: rostamaffendi@utem.edu.my

M. G. Y. Wei · N. S. N. Anwar
Fakulti Kejuruteraan Elektrik, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian, Tunggal, Melaka, Malaysia

Z. Zakaria and R. Ahmad (eds.), *Advances in Electronics Engineering*,
Lecture Notes in Electrical Engineering 619,
https://doi.org/10.1007/978-981-15-1289-6_12

1 Introduction

A depth map is the map produced by stereo matching process. This map contains depth information. This information is useful in many applications such as virtual reality [1], 3D surface reconstruction [2, 3], face recognition [4] and robotics automation [5, 6]. There are short or wide baseline [7] range of applications using the stereo vision sensor. Fundamentally, the triangulation concept is utilized to get the depth estimation based on the pixel information on the depth map. Hence, the matching process is one of the challenging jobs in stereo vision research areas. Basically, matching algorithm consists of multiple stages, which were proposed by Szeliski and Scharstein [8]. The first stage, matching cost computes the preliminary matching point of stereo image. The second stage, the filtering is utilized to reduce the preliminary noise of the first stage. Then, disparity selection and optimization stage normalizes the depth value of each pixel on the image. Last stage is to refine the final result, which is also known as depth map post-processing step.

In stereo matching algorithm development, local [9–11] and global [12] methods are listed as two major approaches in optimizing the depth map. Local approach uses local contents or support windows in computing the depth map. Most of the local-based implementations are using support windows such as fixed window [13, 14], adaptive window [15], convolution neural network [16] and multiple window [17]. Generally, this method applies Winner-Takes-All (WTA) strategy in their third stage of optimization stage [18–20]. Local method has fast running time and low computational complexity. In [21], the RANSAC plane fitting technique was used to increase the efficiency at the final stage of the algorithm development. This method works on low textured regions but unable to correctly determine the edges of object detection. Wrong plane fitting assumptions make incorrect depth estimation. Usually, local-based method shows low precision on the low texture region because there is possibility that improper window sizes were selected during the matching process. Thus, one of the challenges for researches is to overcome this problem.

Fundamentally, global method processes the depth map based on the Markov Random Field (MRF) technique. The MRF uses energy minimization function as implemented in [13] and [22] by using Belief Propagation (BP) and Graph Cut (GC) respectively. These techniques are using energy minimization function for all pixels on an image. It calculates from current point to the nearest pixels of depth map using maximum flow and cuts the minimum energy flow. Base on their methods, the computational requirement is very high, because it counts every pixel individually. Hence, the global method requires long execution time since it uses the iterative technique for each depth estimation. The local-based method is selected in this article. The first stage of the proposed work will be implemented using enhanced modified SAD from [23]. Then the BF is utilized at the second and the last stages where the BF is strong against the low texture regions [24]. The WTA strategy [25] will be executed at the optimization stage. The last stage, MF is utilized to remove remaining noise on the depth map.

2 Methodology

Based on Fig. 1, a flowchart of the proposed work in this article starts with the matching process at STEP 1. This step uses modified SAD to get preliminary result of depth map. The new proposed WSAD based matching cost should be able to increase the effectiveness on the matching process. Then, the BF is used at STEP 2 to filter the noise and preserve the object's edges. After that the WTA strategy will be implemented at STEP 3 by selecting the minimum depth value in an image. The last stage is STEP 4 in the framework, which is using the MF to filter out the remaining noise to obtain the final depth map.

2.1 Matching Cost Computation

This stage produces preliminary differences of depth value. The function used at this stage must be robust and strong against the low texture region. Normally, the mismatch between stereo pair pixels are high at this stage. Hence, this work proposes the weighted SAD, which improves the accuracy on the low texture region. The weighted is imposed to increase the volume number of preliminary absolute differences. The input images are in RGB channels, which the SAD function of the left image I_l and I_r is presented by Eq. (1):

$$WSAD(x, y, d) = \frac{1}{M} \sum_{(x,y) \in M} \left| I_l^i(x, y) - I_r^i(x - d, y) \right| \tag{1}$$

where (x, y, d) represent is the coordinates of depth d, $1/M$ is the weighted over SAD window size, and i denotes the RGB channels of left and right images. Fundamentally, the differences are scaled in pixels-based intensity values.

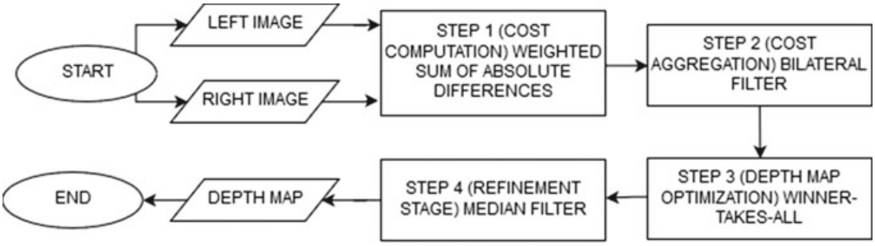

START
LEFT IMAGE
RIGHT IMAGE
STEP 1 (COST COMPUTATION) WEIGHTED SUM OF ABSOLUTE DIFFERENCES
STEP 2 (COST AGGREGATION) BILATERAL FILTER
STEP 3 (DEPTH MAP OPTIMIZATION) WINNER-TAKES-ALL
STEP 4 (REFINEMENT STAGE) MEDIAN FILTER
DEPTH MAP
END

Fig. 1 A flow chart of the proposed algorithm

2.2 Cost Aggregation

The local-based method requires this stage to filter out the preliminary differences after the step of matching cost computation. Hence, this stage is very important to minimize the error due to matching uncertainties on the low texture region. The proposed work at this stage utilizes the BF, which this filter is efficiently removed the noise with, preserved the object edges. The formulation of BF is given by Eq. (2):

$$WM^{BF}_{p,q} = \sum_{q \in w_B} \exp\left(-\frac{|p-q|^2}{\sigma_s^2}\right) exp\left(-\frac{|I_p - I_q|^2}{\sigma_c^2}\right) \tag{2}$$

where q and w_B are the neighbouring pixels and BF support window respectively, p is the positions of pixel (x, y) in the filter windows. The σ_c equals to the color law of similarity factor, and σ_s describes a spatial adjustment factor. The $I_p - I_q$ denotes the Euclidean distance in color space and $p - q$ is the spatial Euclidean interval. The function of this stage is formulated by Eq. (3).

$$C(x, y, d) = WM^{BF}_{x,y,q} WSAD(x, y, d) \tag{3}$$

2.3 Depth Map Optimization

Generally, every image contains a set of depth values. This stage utilizes the WTA strategy, which uses minimum depth value for every location on the depth map. Hence, the WTA is the most suitable approach to be used in this article. The formulation of the WTA is given by Eq. (4):

$$d_{x,y} = \arg\min_{d \in D} C(x, y, d) \tag{4}$$

where C(x, y, d) represents the data of aggregation step and D denotes a set of valid depth values for an image. There are some invalid pixels still remained on the depth map. Hence, this invalid depth will be treated at the next step to increase the accuracy.

2.4 Refinement Stage

This stage consists of several continuous processes. It starts with occlusion handling, invalid pixels filling process, and smoothing the final depth map. The occlusion region comprises the invalid pixels, which is detected by left-right consistency checking process. Then, these invalid pixels will be replaced by a valid pixel value using the fill-in process. Generally, after this process, there are many unwanted pixels or some artifacts on the depth map. Hence, the final step is to smooth the final depth map using Median Filter (MF). The parameters of BP are similarly used as implemented at cost aggregation step. The BF kernel is given by Eq. (5).

$$d_f = median\left\{d_{x,y} | (x, y) \in w_{MF}\right. \tag{5}$$

where d_f is the final depth value at the location (x, y) and w_{MF} represents the MF window size.

3 Results and Discussion

This section explains about the depth map results that will be represented by gray-scale intensity. The darker images show that the respected object is far away from the sensor (i.e., stereo camera). The lighter intensity volume indicates that the object is closer to the sensor. The platform used in this section for experimental analysis is a personal computer with Windows 10, 8G RAM and i5, 3.2 GHz processor. The dataset is using a standard benchmarking evaluation system from the Middlebury [26]. This dataset contains 15 training images with an online submission. The parameters for this article are $\{M,\sigma_s,\sigma_c,w_{B,}w_{MF}\}$ with the values of $\{7 \times 7,17,0.3,13 \times 13,\ 9 \times 9\}$. Figure 2 shows the final results of 15 training images from the Middlebury dataset. The accuracy attributes for error evaluation are *nonocc* and *all error*. The *nonocc error* is the error evaluation based on the non-occluded regions on depth map while *all error* represents the all pixels' evaluation on an image.

Fundamentally, the real images from the Middlebury are difficult and complex to be matched due to different pixel values at the same corresponding point. However, the proposed algorithm in this article has correctly determined the depth location as shown in Fig. 2. Most of the depth levels are assigned at precise positions where the contours of object distance are well-recognized. For example, the last image known as Vintage, wherein the depth of the objects (i.e., computer monitor, CPU, keyboard) are well recovered and reconstructed based on the different depth level. It shows that the proposed work is robust against the input images with different characteristics. Furthermore, based on the experimental results, the depth map was also reconstructed efficiently on the low texture areas (i.e., background and wallpaper of Adirondack image), which increases the accuracy on the depth map result. Figure 2 also shows the final depth images of the Middlebury dataset based on the quantitative results as tabulated in Tables 1 and 2.

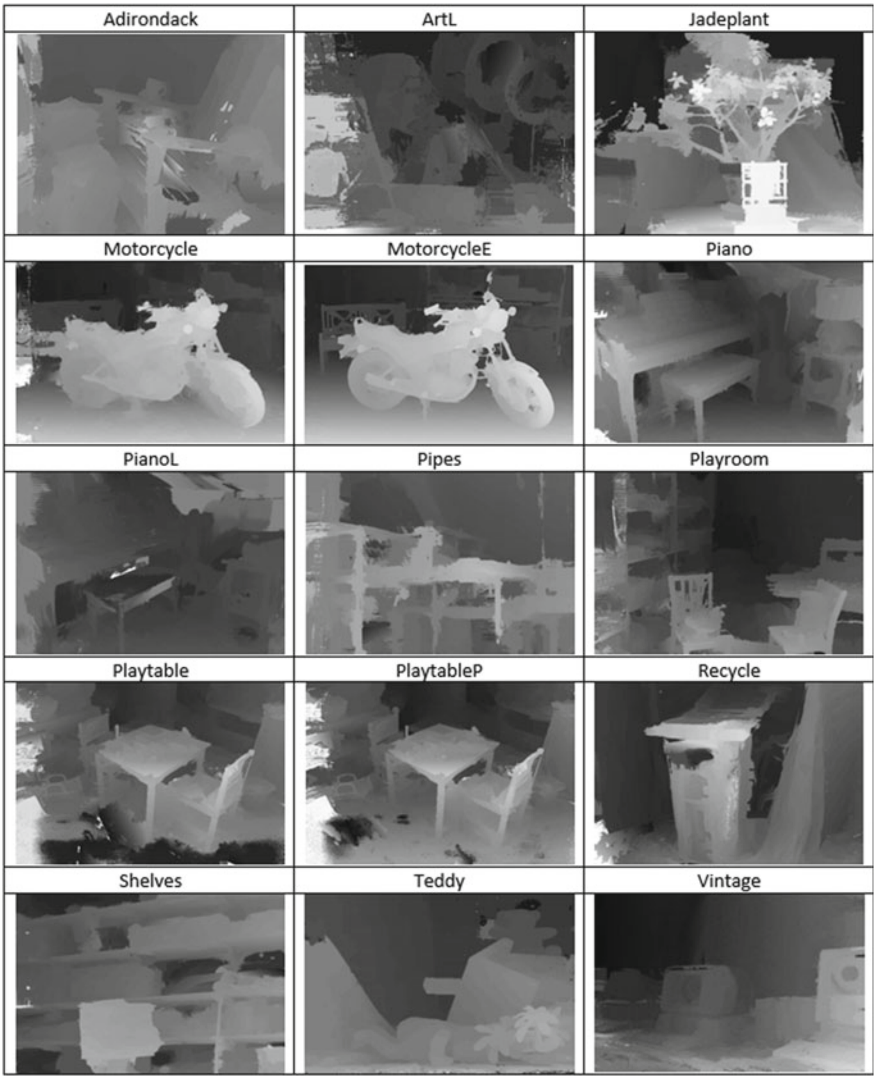


Fig. 2 Depth map of training images from the Middlebury

4 Conclusion

An accurate stereo matching algorithm was presented in this article. The framework used the combination of weighted SAD, using block matching technique based on RGB color differences and gradient matching at the first stage. Then, the second and the last stages utilized an edge-preserving filter, which is able to further reduce the noise based on the standard quantitative benchmarking dataset. The BF used in the

Table 1 The quantitative results of *nonocc* error from the Middlebury

Algorithms	Adiron	ArtL	Jadepl	Motor	MotorE	Piano	PianoL	Pipes	Playrm	Playt	PlayP	Recyc	Shelvs	Teddy	Vintge	Weight Ave
Proposed Algorithm	3.56	4.00	12.12	2.21	2.98	5.74	10.01	6.25	5.87	24.88	5.41	3.62	8.04	2.77	8.21	6.01
SNCC [27]	2.89	4.05	18.10	2.68	2.52	3.52	7.08	6.14	5.64	45.40	3.13	2.90	7.59	1.58	13.50	6.97
ELAS [28]	3.09	4.72	29.70	3.28	3.29	4.30	8.31	5.61	6.00	21.80	2.84	3.09	9.00	2.36	10.90	7.22
MPSV [18]	3.83	6.00	19.70	5.85	5.53	5.68	34.30	9.59	5.86	15.30	4.20	4.59	13.00	3.70	14.30	8.81
ADSM [19]	13.30	6.10	15.00	3.67	5.67	7.08	20.60	6.57	13.20	23.10	3.55	5.76	17.20	3.05	10.10	8.95
DoGGuided [20]	15.20	9.57	27.10	5.64	8.31	8.09	32.40	9.67	14.00	24.50	5.32	5.56	16.20	4.15	15.00	12.00
BSM [29]	7.27	11.40	30.50	6.67	6.52	10.80	32.10	10.50	12.50	24.40	12.80	7.42	16.40	4.88	32.80	13.40

Table 2 The quantitative results of *all* error from the Middlebury

Algorithms	Adiron	ArtL	Jadepl	Motor	MotorE	Piano	PianoL	Pipes	Playrm	Playt	PlayP	Recyc	Shelvs	Teddy	Vintge	Weight Ave
Proposed Algorithm	4.15	7.12	31.97	6.11	4.02	5.18	10.44	12.10	7.15	25.70	6.21	4.44	9.07	3.02	9.56	8.77
SNCC [27]	3.63	6.78	39.80	5.12	5.11	4.65	8.23	11.80	8.05	45.60	4.36	3.29	8.10	2.55	14.80	10.40
ELAS [28]	4.08	7.18	52.80	5.39	5.45	4.96	9.00	10.70	7.94	23.20	3.83	3.78	9.46	3.34	11.60	10.60
ADSM [19]	14.30	10.60	34.10	6.00	8.00	7.37	20.40	12.10	16.90	25.50	5.84	5.83	17.20	4.11	11.10	12.30
MPSV [18]	5.87	9.43	40.20	9.11	8.80	7.03	34.20	15.80	8.58	16.90	5.89	6.78	13.70	4.82	16.80	12.70
DoGGuided [20]	20.10	28.00	56.50	13.80	16.80	13.40	37.30	23.80	30.30	30.80	13.00	9.13	19.00	13.40	23.60	22.30
BSM [29]	12.70	28.70	58.70	14.80	14.70	16.00	35.80	24.50	29.40	31.00	20.20	12.10	19.20	14.30	39.30	23.50

framework increased the accuracy and robust against the different brightness and contrast on the images. Furthermore, the proposed framework is competitive with some established algorithms in the Middlebury database as shown in Tables 1 and 2. It proves that the proposed work in this article can be applied as a complete algorithm in machine vision applications.

Acknowledgements This work was supported by the Universiti Teknikal Malaysia Melaka with the grant Number (PJP/2018/FTK(13C)/S01632).

References

1. Vedamurthy I, Knill DC, Huang SJ, Yung A, Ding J, Kwon OS, Bavelier D, Levi DM (2016) Recovering stereo vision by squashing virtual bugs in a virtual reality environment. Phil Trans R Soc B 371(1697):20150264
2. Hamzah RA, Ibrahim H, Hassan AH (2016) Stereo matching algorithm for 3D surface reconstruction based on triangulation principle. In: International conference on information technology, information systems and electrical engineering (ICITISEE), pp 119–124
3. Menze M, Geiger A (2015) Object scene flow for autonomous vehicles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3061–3070
4. Aziz KA, Shokri AS (2012) A pixel to pixel correspondence and region of interest in stereo vision application. In: 2012 IEEE Symposium on Computers & Informatics (ISCI), pp 193–197
5. Hasan AH, Hamzah RA, Johar MH (2009) Range estimation in disparity mapping for navigation of stereo vision autonomous vehicle using curve fitting tool. IJVIPNS, pp 5–9 (2009)
6. Hamid MS, Rosly HN, Hashim NM (2011) A distance and pixel intensity relation for disparity mapping in region of interest. In: 2011 IEEE 3rd international conference on communication software and networks, pp 15–19
7. Xi HX, Cui W (2013) Wide baseline matching using support vector regression. Telecommun Comput Electron Control 597–602
8. Scharstein D, Szeliski R (2002) A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int J Comput Vis 7–42
9. Yang Q (2012) A non-local cost aggregation method for stereo matching. In IEEE conference on computer vision and pattern recognition (CVPR), pp 1402–14092012)
10. Hosni A, Rhemann C, Bleyer M, Rother C, Gelautz M (2013) Fast cost-volume filtering for visual correspondence and beyond. IEEE Trans Pattern Anal Mach Intell 504–511
11. Kadmin AF, Hamid MS, Ghani SF, Ibrahim H (2018) Improvement of stereo matching algorithm for 3D surface reconstruction. Signal Process: Image Commun, 65:165–172
12. Richardt C, Kim H, Valgaerts L, Theobalt C (2016) Dense wide-baseline scene flow from two handheld video cameras. In: Fourth International Conference on 3D Vision (3DV), pp 276–285
13. Liang Q, Yang Y, Liu B (2014) Stereo matching algorithm based on ground control points using graph cut. Int Congr Image Signal Process (CISP) 503–508
14. Yang Q, Ji P, Li D, Yao S, Zhang M (2014) Fast stereo matching using adaptive guided filtering. Image Vis Comput 202–211
15. Kowalczuk J, Psota ET, Perez LC (2013) Real-time stereo matching on CUDA using an iterative refinement method for adaptive support-weight correspondences. IEEE Trans Circuits Syst Video Technol 94–104

16. Zbontar J, LeCun Y (2015) Computing the stereo matching cost with a convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1592–1599
17. Hirschmüller H, Innocent PR, Garibaldi J (2002) Real-time correlation-based stereo vision with reduced border errors. Int J Comput Vis 229–246
18. Bricola JC, Bilodeau M, Beucher S (2016) Morphological processing of stereoscopic image superimpositions for disparity map estimation. Hal-01330139, pp 1–17
19. Ma N, Men Y, Men C, Li X (2016) Accurate dense stereo matching based on image segmentation using an adaptive multi-cost approach. Symmetry 159
20. Kitagawa M, Shimizu I, Sara R (2017) High accuracy local stereo matching using DoG scale map. In: IAPR international conference on machine vision applications (MVA), pp 258–261
21. Kadmin AF, Ghani SF, Hamid MS, Salam S (2017) Disparity refinement process based on RANSAC plane fitting for machine vision applications. J Fundam Appl Sci 9(4S):226–237
22. Wu SS, Tsai CH, Chen LG (2016) Efficient hardware architecture for large disparity range stereo matching based on belief propagation. In: IEEE international workshop on signal processing systems (SiPS), pp 236–241
23. Hasan AH, Hamzah RA, Johar MH (2009) Disparity mapping for navigation of stereo vision autonomous guided vehicle. In: International conference of soft computing and pattern recognition, pp 575–579
24. Hamzah RA, Rahim RA (2010) Depth evaluation in selected region of disparity mapping for navigation of stereo vision mobile robot. In: IEEE Symposium on Industrial Electronics & Applications (ISIEA), pp. 551–555
25. Ghani SF, Din A, Aziz KA (2012) Visualization of image distortion on camera calibration for stereo vision application. In: International Conference on Control System, Computing and Engineering (ICCSCE), pp. 28–33
26. Daniel S, Richard S (2018) Middlebury stereo evaluation—version 3. Accessed date: Nov 2018, http://vision.middlebury.edu/stereo/eval/references (2018)
27. Einecke N, Eggert J (2013) Anisotropic median filtering for stereo disparity map refinement. In: VISAPP, pp 189–198 (2013)
28. Geiger A, Roser M, Urtasun R (2010) Efficient large-scale stereo matching. In: Asian conference on computer vision, pp 25–38
29. Zhang K, Li J, Li Y, Hu W., Sun L, Yang S (2012) Binary stereo matching. In: International conference on pattern recognition (ICPR), pp 356–359

Pairwise Test Suite Generation Based on Hybrid Artificial Bee Colony Algorithm

Ammar K. Alazzawi, Helmi Md Rais, Shuib Basri and Yazan A. Alsariera

Abstract Software plays an important part of our daily life in order to aid and facilitate our routine tasks, especially a household one. However, the failure of software is a major threat to our lives, particularly the critical applications that employed daily. Due to the large number of inputs as well as time consumption for a test and cost, it is becoming hard to get exhaustive testing for any software in order to fault detection. For this reason, Combinatorial Testing Technique (CTT) is one of the famous techniques that have been used in fault detection of the software systems. Pairwise testing is one of the efficient CTT that used widely for fault detection based on the caused failures by two interactions parameters. There are many researchers that have been developed a pairwise testing strategy. Complementing to the earlier researches, this paper proposes a new pairwise test suite generation called Pairwise Hybrid Artificial Bee Colony (P*h*ABC) strategy based on hybridize of an Artificial Bee Colony (ABC) algorithm with a Particle Swarm Optimization (PSO) algorithm. Empirical results shows that P*h*ABC strategy outperforms other strategies in some cases and provides competitive results in other cases by generating the final test suite.

Keywords Meta-heuristics · Hybrid artificial bee colony · Optimization algorithms · Pairwise testing · Software testing

A. K. Alazzawi (✉) · H. M. Rais · S. Basri
Department of Computer and Information Sciences, Universiti Teknologi PETRONAS,
32610 Bandar Seri Iskandar, Perak, Malaysia
e-mail: ammar_16000020@utp.edu.my

H. M. Rais
e-mail: helmim@utp.edu.my

S. Basri
e-mail: shuib_basri@utp.edu.my

Y. A. Alsariera
Department of Computer Sciences, Northern Border University, Arar 73222, Saudi Arabia
e-mail: yazan.ahmad@nbu.edu.sa

Z. Zakaria and R. Ahmad (eds.), *Advances in Electronics Engineering*,
Lecture Notes in Electrical Engineering 619,
https://doi.org/10.1007/978-981-15-1289-6_13

1 Introduction

Our life increasingly relies on the software along with the advancement of technology such as the communications (e.g. Skype), businesses (e.g. Marketing & Sales), education (e.g. Math Problem Solving), Navigation (e.g. Waze), etc. Software use is unavoidable. Software development leads to creation of a software application with more complexity. Therefore, the expectation of faults that may occurred because of an increase to the very large number of combinations among hundreds of parameters (inputs), will grow [1]. Due to the large number of inputs as well as time consumption for a test and cost, it is becoming hard or impossible to get exhaustive testing for any software in order to fault detection. For this reason, Combinatorial Testing Technique (CTT) is one of the famous techniques that have been used in fault detection of the software systems. A special case of combinatorial testing technique is pairwise testing.

Pairwise testing is one of the efficient and effective Combinatorial Testing Techniques (CTT) for feasible solutions, which is based on the caused failures by two interactions parameters value of the configuration system (inputs) [2, 3]. The main purpose of using pairwise testing strategies because most of the failures as a result of interaction can be caused by two parameters at most depending on investigation [4, 5]. There are many researches have reported that around 75% of errors can be discovered by pairwise testing [6].

Producing the minimum test cases (Which indicates test suite size) that can cover all the parameter’s interaction value is still Nondeterministic Polynomial (NP) hard problem [7–10]. Therefore, CTT strategies are used to decrease the number of test cases and generate the best test suite that can cover all possible combinations. There are many existing strategies in the literature that have been suggested by researchers to decrease the number of test cases and produce the best result that can cover all possible combinations [11–15]. These strategies are designed based on the AI-algorithms to solve the optimization problems in order to produce the optimal solution such as Genetic Algorithm (GA) [16, 17], Harmony Search (HS) [18], Bat Algorithm (BA) [15], Simulated Annealing (SA) [19], Cuckoo Search Algorithm (CS) [20], Ant Colony Algorithm (ACA) [17], Particle Swarm (PS) [21], Artificial Bee Colony (ABC) [22–24], Kindly algorithm (KA) [25], Flower Pollination Algorithm (FPA) [26], Teaching Learning-based Optimization (TLBO) [27], etc. Most of these strategies produced the most optimal test suite size better than other strategies that were based on the greedy, mathematic and random algorithms. Therewith, AI-algorithms need to be run more than one time in order to produce the optimal solution [8, 10].

One of the algorithms that has been proposed recently by Karaboge in 2005 is Artificial Bee Colony (ABC). ABC is a meta-heuristics algorithm that mimics the foraging behaviour of a honeybee within the hive [28]. The ABC executes some certain tasks by the bees, where these bees are divided to three type of bees and everyone has a certain job inside the hive in order to increase the amount of nectar. However, ABC algorithm is similar to other meta-heuristic algorithms that have advantages and disadvantages. Because of a meta-heuristic algorithm's randomization, it is impossible to find an optimization algorithm, which can obtain the global optimum for all optimization problems. Therefore, in this research new optimization algorithm will be adopted called Hybrid Artificial Bee Colony (HABC) algorithm based on hybridize of an Artificial Bee Colony (ABC) algorithm with a Particle Swarm Optimization (PSO) algorithm for pairwise test generation. HABC algorithm mimics the behaviour of honeybee inside the hive, which merges the advantages of ABC algorithm with advantages of PSO algorithm.

The paper is organized as follows: Sect. 2 represents the proposed test case generation algorithm. Section 3 evaluates the P*h*ABC through different benchmark experiments in terms of efficiency and performance. Finally, Sect. 4 concludes the paper and presents our future works.

2 Pairwise Test Suite Generation Algorithm

One of the optimization algorithms that have been proposed recently by Karaboge in 2005 is Artificial Bee Colony (ABC). ABC is a meta-heuristics algorithm that mimics the foraging behaviour of a honeybee within the hive [28]. The ABC executes some certain tasks by the bees, where these bees are divided to three type of bee and everyone has a certain job inside the hive in order to increase the amount of nectar. These bees are employed bees, onlooker bees and scout bees. The first type of bees is employed bee that work to avail the advantage of the explored food sources previously with a higher nectar amount, and exchange the information of food source such as direction, distance and profitability with other types of bees waiting inside the hive (e.g. the number of the food source represents the number of test case). The second type of bees is onlooker bee that will select the food source with higher nectar based on the shared information by the employed bee. The last type of bees is the scout bee that searches the environment randomly in order to detect a new or better than the existing food source. Employed bee represents half of the colony and the other half-represented by onlooker and scout bee, where the number of food source is equal to the number of employed bee.

However, ABC algorithm is similar to other meta-heuristic algorithms that have advantages and disadvantages. Because of a meta-heuristic algorithm's randomization, it is impossible to find an optimization algorithm, which can obtain the global optimum for all optimization problems. One of these algorithms is ABC algorithm, where the weaknesses of ABC are the solution development process, there are insufficiencies because of the simple operation in addition to the speed of convergence of the algorithm is increased [29]. For this reason, the algorithm driven to be stuck in the local optimum for some complex problems as a result of the fast convergence occurred. On the other hand, the information sharing activity of ABC algorithm has been shown a weak performance in the experiments [information-sharing activity is defined by using Eq. (1)] [30]. Several modifications have been done by researchers for ABC algorithm in order to overcome these disadvantages such as change on ABC itself or hybridizing with other algorithms [29, 31].

$$V_{ij} = X_{i,j} + \text{rand}\,[-1, 1](X_{ij} - X_{kj}) \tag{1}$$

In this research, a new Hybrid Artificial Bee Colony (HABC) algorithm have been proposed based on PSO algorithm to overcome the disadvantage of ABC algorithm. The inspiration comes from particle's movement operation of PSO algorithm, where the solution improvement mechanism and information sharing processes are totally different and unique, unlike the information sharing activity that exists in ABC algorithm. Which consists of the unique and special parameter called Weight Factor (w). Velocity parameter function is to the improvement degree based on the previous solution. In addition to the velocity parameter, there is learning factors parameters (C1 and C2), where C1 and C2 function are to determine the relative influence of cognitive (self-confidence) and social (swarm-confidence) components, respectively using Eq. (2).

$$V_{i,d}^{t+1} = W^t * V_{i,d}^t + C_1^t * r_1 * (pbest_{i,d}^t - X_{i,d}^t) + C_2^t * r_2 * (gbest_{i,d}^t - X_{i,d}^t) \tag{2}$$

The movement operation of PSO particles depends on the variable velocity, which is not randomly or arbitrary like that solution improvement in ABC algorithm. The local information comes from the local best solution variable, which interacts with the chosen particle's next move value. The global best solution variable has a great effect on the particle's next move. The ABC algorithms have none of all aforementioned problems. Therefore, this research combined the advantages of ABC with the advantages PSO to overcome the optimization problem. The intelligent behavior of proposed Hybrid Artificial Bee Colony (HABC) can be illustrated as follows:

1. **Initial step**: the initial process of HABC algorithm is the same as the original one in ABC and PSO algorithms, where random search of the environment begins to look for food sources. Producing the initial food sources relies on the range of boundaries for the algorithm's parameters that were defined by using Eq (3).

$$x_{ij} = x_{\min,j} + \text{rand}(0, 1)(x_{\max,j} - x_{\min,j}) \tag{3}$$

2. **Employed bee step**: the number of employed bees is equal to the number of food source (where each employed bee is connected to one food source only). Therefore, the employed bee starts to avail the detected food source and collects the information about the nectar amount. Then, employed bee comes back to the hive to communicate the information of the food source with other bees waiting at the hive in the dance area (Where shares the information by the dancing). After the nectar of food source is exhausted, the employed becomes a scout's bee and starts again to search randomly for a better or new source. Local search

for a new food source is defined by using the local search of PSO in Eq (2) instead of Eq (1). After detecting the food source, the probability of selecting food source is defined by using (4).

$$fitness_i = \begin{cases} \frac{1}{1+f_i}, & if \quad f_{i \geq 0} \\ 1 + |f_i|, & if \quad f_{i<0} \end{cases} \tag{4}$$

3. **Onlooker bee step**: the selection of the food source by the onlooker bee criteria relies on the nectar amounts, where the nectar amounts have been evaluated based on the given information inside the dance area by employed bee. The probability selection of the food source is defined by using Eq. (5).

$$\mathrm{Pi} = \frac{fit_i}{\sum_{n=1}^{sn} fit_n} \tag{5}$$

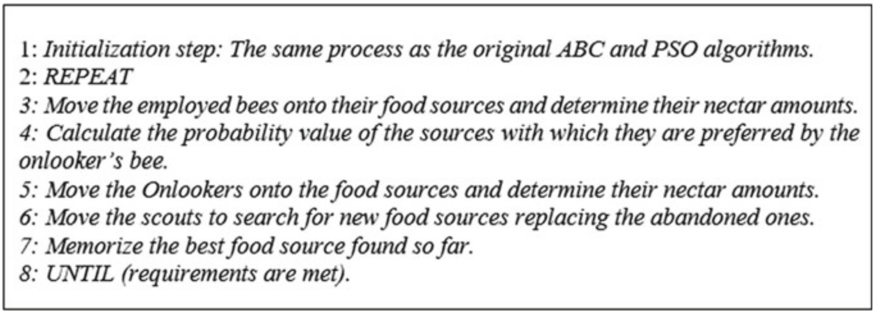
4. **Scout bee and Limit step**: the scout bee mechanism of global search in ABC provides the capability to reduce the convergence problem of early premature. This feature of mechanism is not available in PSO. In addition, during the search of the environment, the algorithm may be trapped in the local minima. For this reason, the "limit parameter" of the ABC is existed in the proposed HABC. The limit parameter of ABC prevents the algorithm from being trapped in the local minima by inserting into the search space a random selected solution from time to time. After the employed and onlooker bees have completed their tasks, the algorithm search the environment in case there is exhausted source to be deserted. The abandoned food source decision relies on counters called limit is defined by using Eq. (6).

$$limit = \mathrm{c.ne.}D \tag{6}$$

During the search, the algorithm will update the counter value and if the counter value is higher than the limit (known as control parameter), then the associated food source with value of the counter will be supposed as abandoned. The discovered new source by scout bee, will replace the abandoned food source. The HABC algorithm main steps are shown as follows in Fig. 1.

3 Result

In this section of the research, the main goal is to benchmark the performance of the proposed P*h*ABC strategy with the existing pairwise testing strategies by using the conducted experiments in [2, 11, 24–26] such as PHSS, Jenny, TVG, IPOG, PABC, PKS, PICT, CTE_XL and TConfig. This research has adopted three experiments to evaluate P*h*ABC strategy with existing pairwise testing strategies. The first



```
Initialization step: The same process as the original ABC and PSO algorithms.
REPEAT
    Move the employed bees onto their food sources and determine their nectar amounts.
    Calculate the probability value of the sources with which they are preferred by the onlooker's bee.
    Move the Onlookers onto the food sources and determine their nectar amounts.
    Move the scouts to search for new food sources replacing the abandoned ones.
    Memorize the best food source found so far.
UNTIL (requirements are met).
```

Fig. 1 HABC algorithm Pseudocode

experiment that has been adopted is CA (N, 2, V^{10}), where *V* represents the variable value from 3 to 10. The second experiment that has been adopted is CA (N, 2, 2^{P}), where *P* represents variable parameters from 3 to 15. The last experiment, comparing P*h*ABC strategy with other existing pairwise testing strategies with 11 configuration system for uniform and non-uniform value.

The P*h*ABC strategy parameters were set at Nbees = 5, maxCycle = 1000, limit = 100, C1 & C2 = 2.0 and W = 0.9. The experiments were implemented twenty independent runs to report the best result due to the randomization characteristic. The experiments were implemented on a Windows 7 (OS) desktop computer with 3.40 GHz Xeon (R) CPU E3 and 8 GB RAM. The Java language JDK 1.8. was used to code and implement the HABC. Tables 1, 2 and 3, present the experimental results and each table presents the optimal test suite size for each configuration. The dark cell with (*) represents the optimal test suite size, while the dark cell without (*) represents the best test suite size shared with other strategies, and the cell with NA represents (not available).

As shown in Table 1, P*h*ABC strategy has produced the most optimal test suite size for CA (N, 2, 5^{10}) with 41 test cases comparing to all other strategies. For CA (N, 2, 3^{10}) and CA (N, 2, 4^{10}), P*h*ABC matches with HSS and PABC by producing 16 and 28 test cases respectively. Whereas HSS produced the most optimal test

Table 1 System Configurations CA (N, 2, V^{10}) where V is ranged from 3 to 10

Value	Tconfig	IPOG	Jenney	CTE_XL	PICT	TVG	PHSS	PABC	PKS	P*h*ABC
3	17	20	19	18	18	18	17	**16**	NA	**16**
4	31	31	30	33	31	33	**28**	30	NA	**28**
5	48	50	45	50	47	50	43	46	NA	**41***
6	64	68	62	71	66	72	**60***	66	NA	67
7	85	90	83	97	88	98	**79***	90	NA	89
8	114	117	**104***	125	112	124	105	118	NA	118
9	139	142	129	161	139	152	**127***	149	NA	149
10	170	176	157	192	170	189	**155***	184	NA	184

Table 2 System Configurations CA (N, 2, 2^P) where P is ranged from 3 to 15

Parameter	Tconfig	IPOG	Jenney	CTE_XL	PICT	TVG	PHSS	PABC	PKS	P*h*ABC
3	**4**	**4**	5	6	**4**	**4**	**4**	**4**	**4**	**4**
4	6	6	6	6	**5**	6	6	**5**	**5**	6
5	**6**	**6**	7	**6**	7	**6**	**6**	**6**	**6**	**6**
6	7	8	8	8	**6**	**6**	7	7	**6**	7
7	9	8	8	8	**7**	8	**7**	**7**	**6**	7
8	9	8	8	8	**7**	8	8	8	**7**	8
9	9	**8**	**8**	9	9	**8**	**8**	**8**	**8**	**8**
10	9	10	10	9	9	9	**8**	**8**	**8**	**8**
11	9	10	9	10	9	9	**8**	9	**8**	**8**
12	9	10	10	10	9	10	9	9	**8**	**8**
13	9	10	10	10	9	10	9	9	**8**	**8**
14	**9**	10	10	10	10	10	10	**9**	**9**	**9**
15	**9**	10	10	10	10	10	10	**9**	**9**	**9**

Table 3 Comparison PhABC with Other Pairwise Strategies for 11 System Configurations

System configurations	Tconfig	IPOG	Jenney	CTE_XL	PICT	TVG	PHSS	PABC	PKS	P*h*ABC
CA (N, 2, 3^3)	10	11	**9**	10	10	11	**9**	**9**	**9**	**9**
CA (N, 2, 3^4)	10	12	13	10	13	12	**9**	**9**	**9**	**9**
CA (N, 2, 3^{13})	20	20	20	21	20	20	**18**	20	20	**18**
CA (N, 2, 10^{10})	170	176	157	192	170	189	**155***	184	184	184
CA (N, 2, 15^{10})	NA	373	**336**	NA	NA	473	341	427	NA	418
CA (N, 2, 10^{20})	NA	NA	NA	NA	NA	NA	**224**	283	NA	278
CA (N, 2, 5^{10})	48	50	45	50	47	50	43	46	46	**40***
MCA (N, 2, 5^1 3^8 2^2)	22	19	41	21	21	23	20	20	20	**15***
MCA (N, 2, 6^1 5^1 4^6 3^8 2^3)	33	36	**31***	39	38	41	39	39	39	36
MCA (N, 2, 7^1 6^1 5^1 4^6 3^8 2^3)	79	44	51	53	46	52	48	47	47	**42***
MCA (N, 2, 10^1 9^1 8^1 7^1 6^1 5^1 4^1 3^1 2^1)	92	**91***	98	102	101	100	95	97	NA	93

cases for the rest of the configurations systems followed by jenny for CA (N, 2, 8^{10}) only. However, P*h*ABC strategy produced competitive and very close results to the optimal result as shown in CA (N, 2, 6^{10}) and CA (N, 5, 7^{10}).

In Table 2, P*h*ABC strategy produced the optimal test cases for the most configuration systems the same as other strategies. Regarding the Table 3, P*h*ABC strategy produced the same test cases exactly like PHSS, PABC and PKS for CA (N, 2, 3^3), CA (N, 2, 3^4) and CA (N, 2, 3^{13}). Whereas P*h*ABC strategy obtains the most optimal test suite size comparing to other strategies for CA (N, 2, 5^{10}), MCA

(N, 2, 6^1 5^1 4^6 3^8 2^3) and MCA (N, 2, 7^1 6^1 5^1 4^6 3^8 2^3) by producing 40, 15 and 42 test cases respectively. However, P*h*ABC strategy produced competitive and very close results to the optimal result for the rest of the configurations systems.

4 Conclusions

This research has proposed P*h*ABC for pairwise test suite generation by relying on hybrid artificial bee colony algorithm. Through the conducted experiments, P*h*ABC has shown a good performance by producing the optimal test suite size for some configuration systems compared with other pairwise strategies. As part of our future research, we are looking to extend the P*h*ABC strategy for supporting high interaction strength to use in future for Software Product Line (SPL) as well as to support the variable strength interaction.

References

1. Yilmaz C, Fouche S, Cohen MB, Porter A, Demiroz G, Koc U (2014) Moving forward with combinatorial interaction testing. Computer 47:37–45
2. Hervieu A, Marijan D, Gotlieb A, Baudry B (2016) Practical minimization of pairwise-covering test configurations using constraint programming. Inf Softw Technol 71:129–146
3. Kuhn DR, Bryce R, Duan F, Ghandehari LS, Lei Y, Kacker RN (2015) Combinatorial testing: theory and practice. Adv Comput 99:1–66
4. Nasser AB, Sariera YA, Alsewari ARA, Zamli KZ (2015) A cuckoo search based pairwise strategy for combinatorial testing problem. J Theor Appl Inf Technol 82:154
5. Kuhn DR, Reilly MJ (2002) An investigation of the applicability of design of experiments to software testing. In: Proceedings of 27th annual NASA goddard/IEEE software engineering workshop. IEEE, pp 91–95
6. Kuhn DR, Wallace DR, Gallo AM (2004) Software fault interactions and implications for software testing. IEEE Trans Software Eng 30:418–421
7. Colbourn CJ, Cohen MB, Turban R (2004) A deterministic density algorithm for pairwise interaction coverage. In: IASTED conference on software engineering, pp 345–352
8. Khalsa SK, Labiche Y (2004) An orchestrated survey of available algorithms and tools for combinatorial testing. In: 2014 IEEE 25th international symposium on software reliability engineering (ISSRE). IEEE, pp 323–334
9. Al-Sewari AA, Zamli KZ (2014) An orchestrated survey on t-way test case generation strategies based on optimization algorithms. In: The 8th international conference on robotic, vision, signal processing & power applications. Springer, pp 255–263
10. Nie C, Leung H (2011) A survey of combinatorial testing. ACM Comput Surv (CSUR) 43:11
11. Alsariera YA, Alamri HS, Zamli KZ (2017) A bat-inspired testing strategy for generating constraints pairwise test suite. In: The 5th international conference on software engineering & computer systems (ICSECS), vol. 5
12. Alsariera YA, Nasser A, Zamli KZ (2016) Benchmarking of Bat-inspired interaction testing strategy. Int J Comput Sci Inf Eng (IJCSIE) 7:71–79
13. Alsariera YA, Zamli KZ (2015) A bat-inspired strategy for t-way interaction testing. Adv Sci Lett 21:2281–2284

14. Alsariera YA, Majid MA, Zamli KZ (2015) Adopting the bat-inspired algorithm for interaction testing. In: The 8th edition of annual conference for software testing, pp 14
15. Alsariera YA, Majid MA, Zamli KZ (2015) A bat-inspired strategy for pairwise testing. ARPN J Eng Appl Sci 10:8500–8506
16. Flores P, Cheon Y (2011) PWiseGen: Generating test cases for pairwise testing using genetic algorithms. In: 2011 IEEE International Conference on Computer Science and Automation Engineering (CSAE). IEEE, pp 747–752
17. Shiba T, Tsuchiya T, Kikuno T (2004) Using artificial life techniques to generate test cases for combinatorial testing. In: Proceedings of the 28th annual international computer software and applications conference, 2004. COMPSAC 2004. IEEE, pp 72–77
18. Alsewari ARA, Zamli KZ (2012) Design and implementation of a harmony-search-based variable-strength t-way testing strategy with constraints support. Inf Softw Technol 54: 553–568
19. Cohen MB, Gibbons PB, Mugridge WB, Colbourn CJ, Collofello JS (2003) A variable strength interaction testing of components. In: Proceedings of 27th annual international computer software and applications conference. COMPSAC 2003. IEEE, pp 413–418
20. Ahmed BS, Abdulsamad TS, Potrus MY (2015) Achievement of minimized combinatorial test suite for configuration-aware software functional testing using the Cuckoo Search algorithm. Inf Softw Technol 66:13–29
21. Rabbi K, Mamun Q, Islam MR (2015) An efficient particle swarm intelligence based strategy to generate optimum test data in t-way testing. In: 2015 IEEE 10th conference on industrial electronics and applications (ICIEA). IEEE, pp. 123–128
22. Alazzawi AK, Rais HM, Basri S (2018) Artificial bee colony algorithm for t-way test suite generation. In: 2018 4th international conference on computer and information sciences (ICCOINS). IEEE, pp. 1–6
23. Alsewari AA, Alazzawi AK, Rassem TH, Kabir MN, Homaid AAB, Alsariera YA, Tairan NM, Zamli KZ (2017) ABC algorithm for combinatorial testing problem. J Telecommun, Electron Comput Eng (JTEC) 9:85–88
24. Alazzawi AK, Homaid AAB, Alomoush AA, Alsewari AA (2017) Artificial bee colony algorithm for pairwise test generation. J Telecommun, Electron Comput Eng (JTEC) 9:103–108
25. Homaid AAB, Alsewari AA, Alazzawi AK, Zamli KZ (2018) A kidney algorithm for pairwise test suite generation. Adv Sci Lett 24:7284–7289
26. Nasser AB, Alsewari AA, Tairan NM, Zamli KZ (2017) Pairwise test data generation based on flower pollination algorithm. Malays J Comput Sci 30:242–257
27. Zamli KZ, Din F, Baharom S, Ahmed BS (2017) Fuzzy adaptive teaching learning-based optimization strategy for the problem of generating mixed strength t-way test suites. Eng Appl Artif Intell 59:35–50
28. Karaboga D (2005) An idea based on honey bee swarm for numerical optimization. Technical report-tr06, Erciyes University, engineering Faculty, Computer Engineering Department
29. Karaboga D, Akay B (2009) A survey: algorithms simulating bee swarm intelligence. Artif Intell Rev 31:61–85
30. Kıran MS, Gündüz M (2012) A novel artificial bee colony-based algorithm for solving the numerical optimization problems. Int J Innov Comput Inf Control 8:6107–6121
31. Yan X, Zhu Y, Zou W (2011) A hybrid artificial bee colony algorithm for numerical function optimization. In: 11th international conference on hybrid Intelligent Systems (HIS), 2011. IEEE, pp 127–132

Language Modelling for a Low-Resource Language in Sarawak, Malaysia

Sarah Samson Juan, Muhamad Fikri Che Ismail, Hamimah Ujir and Irwandi Hipiny

Abstract This paper explores state-of-the-art techniques for creating language models in low-resource setting. It is known that building a good statistical language model requires a large amount of data. Therefore, models that are trained on low-resource language suffer from poor performances. We conducted a study on current language modelling techniques such as *n*-gram and recurrent neural network (RNN) to observe their outcomes on data from a language in Sarawak, Malaysia. The target language is Iban, a widely spoken language in this region. We have collected news data form an online source to build an Iban text corpus. After normalising the data, we trained trigram and RNN language models and tested on automatic speech recognition data. Based on our results, we observed that the RNN language models did not significantly outperform the trigram language models. A slight improvement on RNN model is seen after the size of the training data was increased. We have also experimented on merging *n*-gram and RNN language models and we obtained 32.33% improvement after using a trigram-RNN language model.

Keywords Low-resource language · *n*-gram language model · Recurrent neural network language model

S. S. Juan (✉)
Institute of Social Informatics and Technological Innovations, Universiti Malaysia Sarawak,
94300 Kota Samarahan, Sarawak, Malaysia
e-mail: sjsflora@unimas.my

S. S. Juan · M. F. C. Ismail · H. Ujir · I. Hipiny
Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak,
94300 Kota Samarahan, Sarawak, Malaysia

Z. Zakaria and R. Ahmad (eds.), *Advances in Electronics Engineering*,
Lecture Notes in Electrical Engineering 619,
https://doi.org/10.1007/978-981-15-1289-6_14

1 Introduction

Language modelling is one of the important techniques used in human language technology. The models are used to predict outputs of automatic speech recognition and machine translation. There are several approaches to develop language models such as *n*-gram and recurrent neural network. *n*-gram is a connected sequence of n items from a given text where the items can be phoneme, word or letter [1]. The method has been used for more than 30 years, despite other methods proposed by researchers, due to its capability to compute effectively. However, *n*-gram only depends on a few previous words and has issues when dealing with data sparsity [2]. Recurrent neural network (RNN) [3] on the other hand, is able to overcome these issues and the method has been increasingly used in Natural Language Processing tasks [4–6] due to latest technology available to handle its computational complexity. As both are data-driven methods, typically a big size data is needed to obtain good models to predict word phrases. Therefore, this requirement is a problem when dealing with language with high data sparsity. Languages that have poor orthography system, inconsistency, or lack of digital data could face poor performance in language modelling.

This paper describes state-of-the-art language modelling techniques and their applications to Iban language, a low-resource language in Sarawak, Malaysia. The first part of the paper will explain the theoretical concepts of *n*-gram and RNN techniques followed by a briefed description on low-resource language. Subsequently, we describe our data collection for the experiments which include some statistics of the data. Following that, experiments and results are discussed followed by conclusion and future work of this study.

2 Related Works

2.1 Language Modelling

A language model is a statistical model that can be described as the prior probability of the word sequence. In general, language model is used to predict the probability that a specified word appears next after a given sequence of words [7]. The language model outlines the target language grammar by providing rules for combining words to become expressive phrases as the output of natural language processing applications such as automatic speech recognition system (ASR) [8]. ASR is a system that converts speech to readable texts. It translates speech signals into words or letters [9]. Language model is a component in ASR to search for the most likely word sequence that matches with the signal coming from the acoustic signal analyser. Figure 1 illustrates the general architecture of an ASR and its components. In this section, we discuss the two methods to build language model for ASR, which are *n*-gram and RNN.

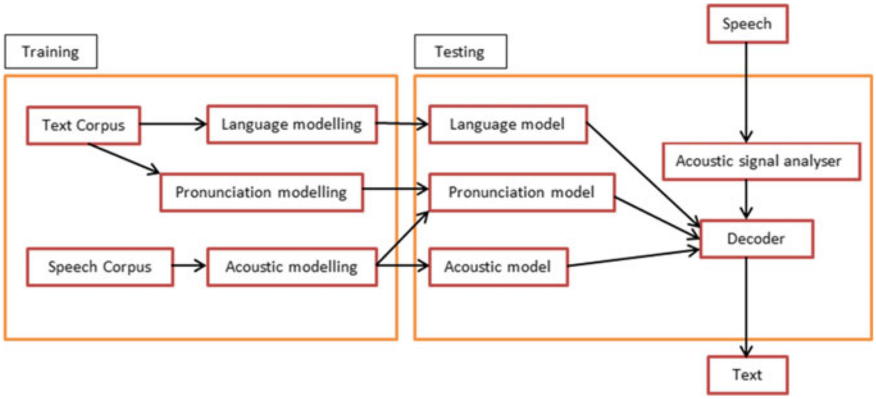


Fig. 1 Schematic view on how ASR system work on transcribing spoken words into written text

n-gram has been used in language modelling for over 20 years. It is able to deal with out-of-grammar utterance and can effectively compute large text corpus. Vu and Schultz [10] describes n-gram as a method that inevitably captures extracted semantic knowledge about the target language from a text corpus and assists in choosing the best option for a word prediction. Typically, a large amount of data is required to calculate the probability of words that appear in a text corpus. n-gram models can be obtained in different order such as unigram, bigram or trigram. The basic idea of n-gram is following a Markov assumption that uses previous words that appear before the target word, history, H; prediction of the occurrence of a word, W can be made [8]. The history consists of the previous n-1 words from a text and depends on order of n where n could be zero (unigram) when no history is account in or, n could be one (bigram) where it contains a context of two words and the history of one word is considered. The probability of a sentence $P(W)$ to have series of $W = w_1, w_2, w_3, \ldots w_m$ words can be denoted as

$$\begin{aligned} P(W) &= P(w_1, w_2, w_3, \ldots w_m) \\ &= P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)\ldots P(w_m|w_1, \ldots, w_{m-1}) \end{aligned} \tag{1}$$

The chain rule is applied to decompose the joint probability into a set of conditional probabilities where the product is:

$$\prod_{m}^{n} P(w_m|w_1, w_2, \ldots, w_{m-1}) \tag{2}$$

To explain how the word probability can be calculated, we use a simple example. If we have a sentence such as, “I love cake”. The unigram of the sentence can be represented as P(I), P(love) and P(cake). To determine the probability of obtaining the word “love” after “I”, thus the bigram is computed as:

$$P(\text{love}|\text{I}) = P(w_{m-1}|w_{m-2}) \tag{3}$$

This can be followed by trigram and since we know that the word "cake" appears after the phrase "I love", therefore the it is calculated as:

$$P(\text{cake }|\text{love I}) = P(w_m|w_{m-1,}w_{m-2}) \tag{4}$$

From Eqs. 3 and 4, we can arrange the probability of the sentence as follows:

$$P(\text{I love cake}) = P(\text{I}) \times P(\text{love}|\text{I}) \times P(\text{cake}|\text{I love}) \tag{5}$$

Equation 5 shows a trigram model that calculates the probability occurrence of a collection of three words that appears at the same time. The equation is then could be deduced into:

$$P(w_m|w_{m-1}w_{m-2}) = \frac{C(w_{n-2}, w_{n-1}, w_n)}{C(w_{n-2}, w_{n-1})} \tag{6}$$

where C is defined as the n-gram count. The drawback of this model on estimating word sequence in a data set is that some n-grams might have a very small count or does not appear at all. This phenomenon is known as "unseen n-grams" and this occurs in data sparsity problem.

RNN is one of advanced techniques in language modelling, as presented by Mikolov et al. [3]. From the experiments conducted by the authors, it is shown that RNN language models (RNNLM) were able to obtain 50% reduction on perplexity and reduce word error rate of automatic speech recognition. RNNLM surpassed n-gram model performance based on its ability to use larger size of word context that allows information to cycle inside the network for a long time. RNNLM method exploits the architecture introduced by Jeffery L. Elman that is known as the Elman network [11]. The neural network is a modest network available and much easier to employ and train compared to other neural network [3]. A schematic diagram of the RNNLM architecture is presented in Fig. 2. In RNNLM, there are three layers namely, input layer, hidden layer or context layer, and output layer.

Input into the network in time t is labelled as $x(t)$, state of the network or hidden layer as $s_j(t)$ with j number of layers and the output as $y_k(t)$. The input layer $x(t)$ is determined by concatenating context layer s at time $t-1$, with vector w which signifies current word. $f(z)$ is a sigmoid activation function and can be denoted as;

$$f(z) = \frac{1}{1+e^{-z}} \tag{7}$$

while $g(z)$ is the softmax function;

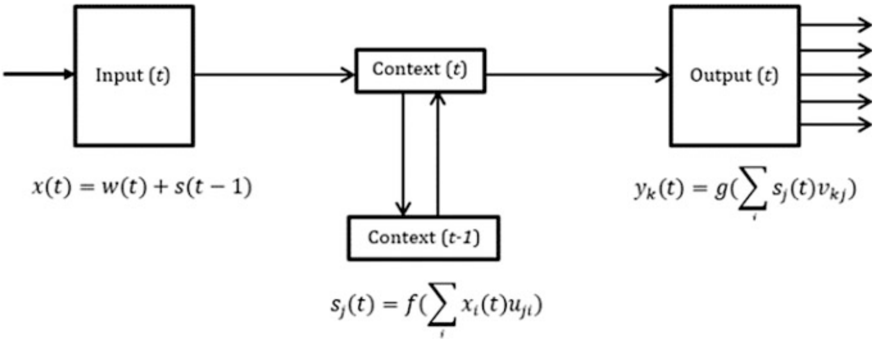


Fig. 2 Schematic view of simple recurrent neural network or Elman network exploited for generating language model [3]

$$g(z_m) = \frac{e^{z_m}}{\sum_k e^{z_k}} \tag{8}$$

Output layer $y(t)$ indicates probability distribution of next word based on previous word $w(t)$ and context layer $s(t-1)$ while the softmax function certifies this probability distribution is valid.

RNNLM has been reviewed to have a few limitations such as longer training time, fixed number of hidden neurons and the length of context to use is limited (De Mulder et al. 2015). Tests conducted by Mikolov et al. [3] on New York Time (NYT) section of English Gigaword took several weeks of training. This training complexity is related to the size of vocabulary used as well as the parameters to obtain improvements in accuracy. Besides that, determining the number of hidden neurons is also an issue for RNNLM. To train an RNNLM, one may need to set a random number of hidden neurons at the beginning and then, fine-tune by changing parameters to get better results.

2.2 *Evaluating Language Model*

The performance of a language models can be measured through its level of perplexity [7]. Perplexity is defined as the inverse of geometric average probability given by the model to every word in a training data. This technique is popular for testing language models as it makes easy comparison between each model's performances. A good language model is determined by obtaining a low perplexity value. Perplexity is computed as 2^H where

$$H = -\frac{1}{N}\sum_{i=1}^{n}\log_2 P(s_i) \tag{9}$$

and N is the total number of words in a test corpus, while $\sum_{i=1}^{n}\log_2 P(s_i)$ measures the uncertainty of the corpus that has n number of sentences.

2.3 Low-Resource Language in Malaysia

In human language technology (HLT), a low-resource or under-resourced language is described as a language that has unwritten or limited resource for orthography system, inadequate occurrence on the web, low number of language experts or lack of automated resources for speech or text processing such as pronunciation dictionary or part of speech [12, 13]. Malaysia, for instance, has 140 living languages that are mostly unwritten, thus, there are many low-resource languages in this country. A huge challenge for researchers or developers to develop HLT applications such as automatic speech recognition, machine translation, speech synthesis; due to lack of training data to develop statistical models.

Iban is a language spoken in the Borneo island, primarily in Sarawak, Kalimantan and Brunei [14]. According to Ethnologue, Iban belongs to the Austronesian family tree in the branch of Malayo-Polynesian, Malayo-Chamic, Malayic, and Ibanic and it is spoken by 752,000 speakers [15]. The language is taught in primary and secondary schools in the Sarawak state and it is also offered as an undergraduate course in several local universities. The Borneo Post, a local newspaper in arawak provides news in Iban and publishes articles both in print and online.[1] There are also other websites that regularly publish articles in Iban such as Pegari[2] and Dayak Daily.[3] Research in HLT applications for Iban is still poor due to lack of digital resources that are available, hence, Iban is considered a low-resource language.

3 Iban Language Data

For language modelling experiments, we used two datasets; an existing Iban corpus that was obtained from an online Iban news collected from 2007 to 2012 for a previous work [13] and a recently acquired Iban news data collected from 2012 to

[1] https://www.utusanborneo.com.my/iban.

[2] https://pegari02.wordpress.com/servis/majalah-pegari/.

[3] https://dayakdaily.com/category/iban-section/.

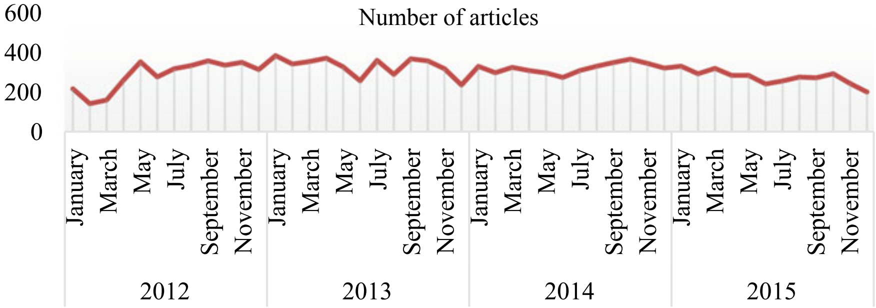


Fig. 3 Total number of Iban articles from 2012 to 2015

2015. The 4-year corpus has a total of 14,533 Iban articles from 1st January 2012 until 31st December 2015. Figure 3 shows the amount of Iban articles produced by the news agency each month during the mentioned period.

From the graph, we can observe that the lowest number of articles recorded is in February 2012 with 142 articles while January 2013 makes the highest count with 384 articles. The high number of articles produced in early 2013 was the effect from the 13th Malaysia's General Election where more political news was published by the agency. We also observe that, there is a noticeable decrease on articles published in both June and December each year due to long holidays during Gawai Dayak festival and Christmas.

3.1 Normalizing the Text Corpus

In Natural Language Processing (NLP), data normalization is performed to convert desired corpus to a more convenient and standard form [8]. It is a process of converting numbers, dates, acronyms, and abbreviation that is pronounced directly in its context. Elements such as punctuations and numbers can influence performance of language models as it will consider the presence of these elements as tokens. The steps for normalizing Iban data are as follows:

- Removing acronyms in brackets—Acronyms that appear after its original meaning is removed as it is considered as duplication. For example, *Universiti Malaysia Sarawak* (UNIMAS).
- Converting entire corpus to UTF8 encoding—This is done due to the fact that some of the articles may contain punctuation in different encoding.
- Removing punctuation—Punctuation such as question marks, exclamation marks, full stops and comma to name a few, are removed as presence of these element will affect the accountability of language models. However, we keep hyphenation '-'in the text, as it is frequently used in reduplication such as '*tajau-tajau*' and '*chamang-chamang*'.

- Changing selected acronyms to its actual meaning—It is common to use acronyms or abbreviation in writing such as ‘*Doktor*’ which is used as ‘Dr’.
- Changing numbers to words—This is done to convert number such as ‘1’ and ‘234’ to Iban words which is ‘*satu*’ and ‘*dua ratus tiga puluh empat*’, respectively.
- Changing all upper-case letter to lower case letter.

After normalizing the 2012–2015 articles, our corpus contains 307 K characters, 18 K lines and 48 K words.

4 Iban Language Modelling Experiments

To train and test our models, we have utilized open source tools namely, SRI [16] and RNNLM language modelling toolkits [17].

4.1 Building n-Gram Language Models

We trained trigram language models and applied two smoothing methods, Kneser-Ney [18] and Witten-Bell [19] or KN and WB, respectively. Three datasets were used to train the models. The first set contains 2007–2012 Iban news articles, the second set has 2012–2015 news articles and the third one is a combination of the first two sets. Then, the language models were evaluated using a test corpus that was used for automatic speech recognition experiments [20]. Results of the *n*-gram language modelling experiment are tabulated in Table 1.

Based on the results, it is observed that the language models that were trained on the newly acquired data have the highest perplexity values. Combining the two data for training language models gave us lower perplexity values, however, the differences are slightly significant. The language model trained on combined data (2007–2015) only gained 4% (KN) and 3% (WB) percent improvement from the language model trained on 2007–2012 data. Besides that, Kneser-Ney smoothing outperformed Witten-Bell smoothing on all datasets.

Table 1 Results of the *n*-gram language models which were obtained using three Iban corpora

Corpus	Language model	Perplexity
2007–2012	3-gram-KN	159
	3-gram-WB	166
2012–2015	3-gram-KN	186
	3-gram-WB	197
2007–2015 (combine)	3-gram-KN	152
	3-gram-WB	161

4.2 Obtaining RNN Language Models

To get baseline RNNLMs, we used 40 hidden layers and Table 2 shows results obtained from the experiment.

Both language models have perplexity of 222 and 359, respectively, which did not outperform the *n*-gram models. We attempt to improve this baseline by applying stochastic gradient descent Back Propagate Through Time (BPTT) in training RNNLM. As shown in Table 3, we gained 7.66% and 26.86% improvements from the baseline models for models trained using 2007–2012 corpus and 2012–2015 corpus, respectively. We also combined the two corpora to obtain another RNNLM and it has shown that RNNLM with 30 layers and using BPPT outperformed the two previous RNNLMs.

To investigate the effects of different number of hidden layers, we trained RNNLMs using 100 hidden layers and applied the BPPT algorithm in the training. Our results are shown in Table 4.

We obtained better perplexity value for RNNLMs that were trained on 2012–2015 corpus and 2007–2015 corpus. We achieved 34.36% improvement from the baseline of RNNLM trained on 2012–2015 corpus and 5.68% improvement from the RNNLM trained using 30 hidden layers on 2007–2015 corpus. Although the latter RNNLM performs the best, the training time to obtain the model was the longest (1.6 h) due to the size of the layers and corpus. On the other hand, there is not much difference in terms of the results of RNNLMs with 40 and 100 layers, which we obtained using 2007–2012 corpus.

Table 2 Baseline results of the RNN language models using RNN language modeling toolkit

Corpus	Language model	Perplexity
2007–2012	lm0712_base	222
2012–2015	lm1215_base	359

Table 3 Results of the RNN language models after applying BPPT

Corpus	Language model	Perplexity
2007–2012	lm0712-40-bppt	205
2012–2015	lm1215-40-bppt	283
2007–2015	lm0715-30-bppt	186

Table 4 Results of RNN language models with 100 hidden layers

Corpus	Language model	Perplexity
2007–2012	lm0712-100	208
2012–2015	lm1215-100	267
2007–2015	lm0715-100	176

Table 5 Results of Iban *n*-gram-RNN language models

Language model	Perplexity
lm3gram0712 + lm0712-40	141
lm3gram1215 + lm1215-100	179
lm3gram0715 + lm0715-100	133

4.3 Merging n-Gram and RNNLM Language Models

Researchers [17] has shown that combining *n*-gram and RNN language models results a significant reduction of perplexity value. Instead of adding new data for training, the approach uses existing models and merge them to get better results. We applied this strategy to observe the impact to our current results. In this experiment, we merged our best trigram language models with RNNLMs using the RNN language modelling toolkit. On the 2007–2012 corpus, RNNLMs were trained with 40 hidden layers and BPPT applied. For 2012–2015 and concatenated corpora, we used RNNLMs with 100 layers and BPPT applied. Results of the merging models experiment are tabulated in Table 5.

By applying this approach, our best perplexity value is obtained after merging the *n*-gram and RNN models trained from the largest corpus. The result has also outperformed results from our previous experiments where we gained 32.33% improvement from the RNNLM trained using 2007–2015 corpus. Hence, this proves that merging models can effectively improve our language model's performance. We also observe that the size of the corpus still effects the perplexity of the language model.

5 Conclusions and Future Work

In this paper, we have demonstrated our attempt in building language models using conventional and advanced statistical techniques for Iban language, a low-resource language. Online news articles were collected from a single source to build text corpus for training language models. An existing corpus was also used in our experiments to compare results. All our models were trained on three datasets, the 2012–2015 corpus, 2007–2012 corpus and 2007–2015 corpus, where the latter was obtained by combining the first two datasets. We built trigram models with smoothing methods applied and recurrent neural network (RNN) language models of 40 and 100 layers, with and without back propagation through time (BPPT). Based on the results of our experiments, trigram language models with Kneser-Ney smoothing gave better perplexities and the trigram language models trained on the combined dataset (2007–2015) have the best results. Our RNNLMs performed better when we use larger number of hidden layers and BPPT. Like the *n*-gram experiment results, RNNLM trained on the combined dataset gave the best perplexity value. Thus, we can conclude that larger corpus can improve the *n*-gram and

RNNLM, respectively. However, the performance of our RNNLMs did not outperform the trigram models. On the other hand, merging RNNLM and n-gram showed us significant improvements, whereby the lowest perplexity we obtained is 133. This result was obtained from the merging models that were trained on the combined dataset. It shows that merging models that are obtained from n-gram and RNN methods was able to improve our Iban language model. In our future work, we plan to test the n-gram-RNNLM language model to predict outputs of an Iban automatic speech recognition system.

References

1. Koehn P (2009) Statistical machine translation. Cambridge University Press, Cambridge
2. Chen X, Liu X, Qian Y, Gales MJF, Woodland PC (2016) CUED-RNNLM—an open-source toolkit for efficient training and evaluation of recurrent neural network language models. ICASSP, IEEE international conference on acoustics, speech and signal processing, vol 2016 May, pp 6000–6004
3. Mikolov T, Karafiát M, Burget L, Cernocký J, Khudanpur S (2010) Recurrent neural network language modeling. September, pp 1045–1048
4. Chen Z-M, Guo X-Q, Huang Y-F, Zhang Y-J, Yang Z-L (2018) RNN-stega: linguistic steganography based on recurrent neural networks. IEEE Trans Inf Forensics Secur 14 (5):1280–1295
5. Park H, Cho S, Park J (2018) Word RNN as a baseline for sentence completion. In: 2018 IEEE 5th international congress on information sciences and technology, pp 183–187
6. Doval Y, Gómez-Rodríguez C (2019) Comparing neural- and N-gram-based language models for word segmentation. J Assoc Inf Sci Technol 70(2):187–197
7. De Mulder W, Bethard S, Moens MF (2015) A survey on the application of recurrent neural networks to statistical language modeling. Comput Speech Lang 30(1):61–98
8. Jurafsky D, Martin JH (2008) Speech and language processing, vol 1
9. Glass J (2007) A brief introduction to automatic speech recognition. Artif Intell 1–22
10. Vu NT, Schultz T (2014) Automatic speech recognition for low-resource languages and accents using multilingual and crosslingual information, p 181
11. Elman JL (1990) Finding structure in time. Cogn Sci
12. Besacier L, Barnard E, Karpov A, Schultz T (2014) Automatic speech recognition for under-resourced languages: a survey. Speech Commun
13. Juan SS (2015) Exploiting resources from closely-related languages for automatic speech recognition in low-resource languages from Malaysia (Phd thesis)
14. Omar A (1981) Phonology. Dewan Bahasa dan Pustaka, Kuala Lumpur
15. Simons GF, Fennig CD (2018) Ethnologue: languages of the world, Twenty-Fir. Dallas, SIL International, Texas
16. Stolcke A (2002) SRILM—an extensible language modeling toolkit. In: 7th international conference on spoken language processing. ICSLP2002—INTERSPEECH 2002, Denver, Color. USA, Sept. 16–20, 2002, vol 2, no. Denver, Colorado, pp 901–904
17. Mikolov T, Kombrink S, Deoras A, Burget L, Cernocký J (2011) RNNLM—recurrent neural network language modeling toolkit. In: IEEE automatic speech recognition and understanding workshop
18. James F (2000) Modified Kneser-Ney smoothing of n-gram models modified kneser-ney smoothing of n-gram models

19. Witten IH, Bell TC (1991) The zero-frequency problem: estimating the probabilities of novel events in adaptive text compression. IEEE Trans Inf Theory
20. Juan SS, Besacier L, Lecouteux B, Dyab M (2015) Using resources from a closely-related language to develop ASR for a very under-resourced language: a case study for iban. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, 2015, vol 2015 January

Deep Convolutional Network Approach in Spike Train Analysis of Physiotherapy Movements

Fadilla Atyka Nor Rashid, Nor Surayahani Suriani, Mohd Norzali Mohd, Mohd Razali Tomari, Wan Nurshazwani Wan Zakaria and Ain Nazari

Abstract Classifying gestures or movements nowadays have become a demanding business as the technologies of sensors have risen. This has enchanted many researchers to actively and widely investigate within the area of computer vision. Physiotherapy is an action or movement in restoring someone's to health where they need continuous sessions for a period of time in order to gain back the ability to cope with daily living tasks. The rehabilitation sessions basically need to be monitored as it is essential to not just keep on track with the patients' progression, but as well as verifying the correctness of the exercises being performed by the patients. Therefore, this research intended to classify different types of exercises by implementing spike train features into deep learning. This work adopted a dataset from UI-PRMD that was assembled from 10 rehabilitation movements. The data has been encoded into spike trains for spike patterns analysis. Spike train is the foremost choice as features that are hugely rewarding towards deep learning as they can visually differentiate each of the physiotherapy movements with their unique patterns. Deep Convolutional Network then takes place for classification to improve the validity and robustness of the whole model. The result found that the proposed model achieved 0.77 accuracy, which presumed to be a better result in the future.

Keywords Convolutional neural network · Spike train · Rehabilitation · Deep learning

F. A. N. Rashid · N. S. Suriani (✉) · M. N. Mohd · M. R. Tomari ·
W. N. W. Zakaria · A. Nazari
Department of Electronic Engineering, Faculty of Electrical and Electronic Engineering,
Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia
e-mail: nsraya@uthm.edu.my

Z. Zakaria and R. Ahmad (eds.), *Advances in Electronics Engineering*,
Lecture Notes in Electrical Engineering 619,
https://doi.org/10.1007/978-981-15-1289-6_15

1 Introduction

Physiotherapy, also known as physical therapy, is an ongoing treatment for those who struggle with serious illnesses, such as stroke, Parkinson's, post-surgeries, etc. This therapy is extremely helpful in order for the patients to cope with their daily living tasks due to their movement's impairment. The assessment movements usually will be prescribed according to the patient's condition within their own pace as the movement for a patient may not be equally adequate for others [1]. Furthermore, the patients need to be monitored while performing the movements so that they execute correct movements and their progression can be tracked by medical experts. However, lack of experts makes physiotherapy session to be delayed and cause discomfort to the patients and the caregivers as well. Hence, technologies take places in assisting the experts conducting the session. Physiotherapy has caught interests of many researchers in regard to machine learning approaches. Different machine learning algorithms have been utilized for recognizing different types of physiotherapy movements, also recognizing different parts of the body.

Recently, neural networks architecture has been a favour for its flexibility and compatibility in a wide range of machine learning applications, including computer vision [2]. In the last few decades, the 3rd generation of neural networks, Spiking Neural Networks (SNNs) has been introduced. SNNs are said to be a promising paradigm [3] as they are practically efficient for modelling complex information, and they are adequate to represent and integrate different information dimensions (time, space, frequency, phase). Their flexibility and self-organising manner make them effortless, when they are dealing with large volumes of data and using information representation as trains of spikes. Thus, SNNs intent to span the gap between neuroscience and machine learning, which leads to boosting a number of researches with the focus on biologically motivated approaches for pattern recognition. This paper employs University of Idaho—Physical Rehabilitation Movements Data Set (UI-PRMD) which is a data set of movements performed by patients in physical therapy and rehabilitation programs to classify the types of the movements. This paper proposed spike trains as feature analysis, which is highly informative for action recognition. Next, the analysis will be undergone through deep learning to classify the types of the movements. This paper is structured as follows: Sect. 2 summarizes related works in classification using UI-PRMD. Section 3 explains the proposed methodology in analysing spike train and classifying by deep learning. Section 4 presents experimental results and discussions of the proposed approaches. Finally, conclusions are drawn in Sect. 5.

2 Previous Work

Machine learning has been said to be potentially guiding rehabilitation sessions, especially in-home rehab sessions as it can improve the ability of recognition and classification. These techniques nowadays are being used extensively in biomedical applications [4], for example in recognising body parts and the types of movements. Previously, Zhu et al. [5] applied K-Nearest Neighbors (KNN) algorithm on rehabilitation prediction, resulting that KNN made significantly better prediction that the clinical assessment protocol, ADLCAP, which was used within the health assessment information systems in Canada. Next, they employed Support Vector Machine in [6] to improve upon KNN algorithms as SVM may give more accurate predictions. However, it turns out that SVM does not statistically improve over KNN. Muniz et al. [7] compared three different models, SVM, Probabilistic Neural Network (PNN), and Logistic Regression (LR) for discriminating between normal and Parkinson diseased people by monitoring their walking postures. PNN seems to be performing better than SVM and LR, as PNN showed high performance indexes in classifying ground reaction force of normal and Parkinson subjects.

While Patsadu et al. [8] compared four approaches for pose classification: Backpropagation Neural Network (BNN), SVM, Decision Tree, and Naïve Bayes. This resulted in BNN and SVM protrude over the other approaches. On the contrary, Neural Networks (NN) and Support Vector Machine (SVM) are also commonly used for posing and motion recognition. Suriani et al. [9] employs SVM to identify a person's state, whether they are in a normal or anomaly movement for fall detection, while doing home-based rehabilitation exercises. On the other hand, there are limitations in speed and size for SVM classification. Ciresan et al. used Convolutional Neural Networks (CNN) [10], while Toshev et al. implemented Deep Neural Networks (DNN) to recognize human poses and activities [11]. Du et al. divided human skeleton into five segments and used each of the parts to train a hierarchical recurrent neural network [12].

DNNs are said to be historically brain-inspired, but the fundamental is different in structure, neural computations, and learning rules compare to the brain. Neurons communicate to each other in a neural circuit by sending spike trains, which sparse in time so that they have high information content. Presently, artificial sensors do not generate spikes as the primary signal, instead they are providing us with an electrical analogue or digital output encoding its measured value. Thus, various spike encoding methods are being introduced [13]. Spike train features have been widely implemented in various areas such as speech recognition technology, as speech signals were converted into spike trains signatures, where they are able to differentiate the speech signals that represent different words [14]. In [15], they equipped spike trains for recognition of object and handwritten characters, classified with SVM. Next, [16, 17] converted analogue signals of SA-I mechanoreceptors in human skin into spike trains by an Izhikevich'model and classified naturalistic textures with an accuracy of 97%.

However, CNN are more widely used types of deep artificial neural networks in diverse fields, such as video and image recognition, and natural language processing along with speech processing. CNN has shown a promising performance in number of image recognition tasks, as the architecture layer, structured by node in the mth layer, being connected to n nodes in the (m − 1) th layer, where n is the size of the receptive field of the CNN. This reduces the total number of parameters in the network and prevents overfitting and ensures a built-in invariance. Nevertheless, researches recently focused towards Deep Learning (DL), where the architecture uses many layers of trainable parameters and has demonstrated outstanding performance in machine learning and AI applications. Therefore, this study proposed to take advantage of spike train ability into a DL framework to develop an encouraging architecture to achieve high performance of proved deep networks, while implementing bio-inspired, power-efficient platforms.

3 Methodology

3.1 Spike Train

Spikes or also known as neuronal action potentials are basically a language that neurons use to delegate and convey information. This neuron model propagates the sequences of spikes, which transmit information to each other in a membrane potential. The sequences of spikes, or also to be called spike trains, consisting of n spikes occurring at time t_i, can be represented mathematically as

$$\rho(t) = \sum_{i=1}^{n} \delta(t - t_{\mathrm{i}}) \tag{1}$$

where δ is the Dirac delta function. In order to quantify the average responses encoded by a single spike train, all that needs to done is to simply add up the spikes over some time interval T:

$$\frac{1}{T}\int_{0}^{T} \rho(\tau)d\tau \tag{2}$$

This is occasionally called the “firing rate” or “spike count rate”. Next, we define an average firing rate over the trials as below:

$$\frac{1}{T}\int_{0}^{T} \langle\rho(\tau)\rangle d\tau \tag{3}$$

where $\langle\,\rangle$ denotes the average over trials. We also define a firing rate with finer temporal resolution for defining a firing rate $r(t)$ as

$$r(t) = \frac{1}{\Delta t}\int\limits_{t}^{t+\Delta t} \langle \rho(\tau)\rangle d\tau \tag{4}$$

Which also can be used to define Post-Stimulus Time Histogram (PSTH). However, PSTH is always defined by depending on where the time bins are placed. Hence, another way to define a firing rate is

$$r(t) = \int\limits_{-\infty}^{\infty} \omega(\tau)\rho(t-\tau)d\tau \tag{5}$$

where $\omega(\tau)$ is called the filter kernel. Nonetheless, a neuron does not have access to the latter information as it is not causal. Thus, a better model conjoint with causal filter is:

$$\omega(t) = \left[\alpha^2\tau\exp(-\alpha\tau)\right]_+ \tag{6}$$

where $[x]_+$ means x for $x \geq 0$, and 0, otherwise. Consequently, we can fit the firing rate to a simple function of the stimulus attributes.

3.2 Convolutional Neural Network (CNN)

Convolutional Neural Networks (CovNets or CNNs) are proclaimed to be a type of Neural Networks that are very effective in areas such as image recognition and classification. For an example, CNNs have been successful in the ability of suggesting with relevant captions for a specific image other than recognizing everyday objects, humans, and animals. CNNs also empower the vision in robots and self-driving cars. CNNs have a bit different architecture than regular Neural Networks. The layers in CNNs are organized in 3 dimensions: width, height, and depth. In addition, the neurons in a layer only associate with a small part of neurons instead of all of them. Hence, the final output will be reduced to a single vector of probability scores, formed along the depth dimension.

There are various influential architectures of CNNs. The very first CNN was LeNet introduced by LeCun [18] in 1998. After over 10 years CNN being incubated, as more data and computing power became available, CNN became more interesting and many architectures of CNNs have been released along the years, such as AlexNet [2], ZF Net [19], GoogLeNet [11], VGGNet [20], ResNets [21], and DenseNet [22]. However, as of May 2016, ResNet, or Residual Network are the

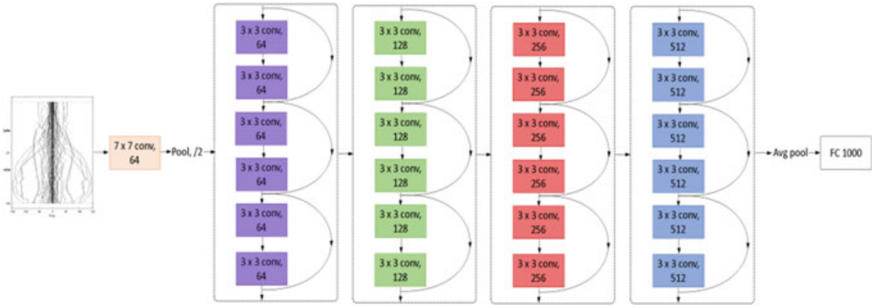


Fig. 1 Architecture of proposed method with ResNet-50 implementation

state-of-the-art CNN models and the default choice for using CNNs in practice. Therefore, this paper chose to adopt ResNet as our CNN model for classification part as it is possible to train large size of layers and achieving fascinating results. As illustrated in Fig. 1, we are having the spike train as the network inputs (798 × 720 × 3 image) and adopting ResNet-50 [21]. The layers are 3 × 3 convolutions and subsampling and they are performed by convolutions with a stride of 2. The networks end with a global average pooling a 10-way fully connected layer and softmax.

4 Result and Analysis

4.1 Experimental Setting

This study proposed an algorithm that adopts spike train features into deep learning approach, which is CNN classification model. The experiment was carried out using normal CPU and the computational time was only 30 s for 100 samples. Figure 2 illustrated the general framework of proposed algorithm. Raw data was collected from several chosen exercises, then it was parted into 100 frames for each sample and generated into spike trains. Hence, each exercise consists of approximately 100 parts of trains.

We then implemented CNN classification model movement identification and categorization. CNN has learnt the spike patterns; thus, it can classify the exercises based on the uniqueness. Ultimately, the results have been tabulated with confusion matrix for validation and summarization of the prediction. Confusion matrix is often used to describe the performance of an algorithm of a classification in visualization, which eases the identification of confusion between classes. Further discussion of each part of the framework has been presented in detail in each of the subtopics below.

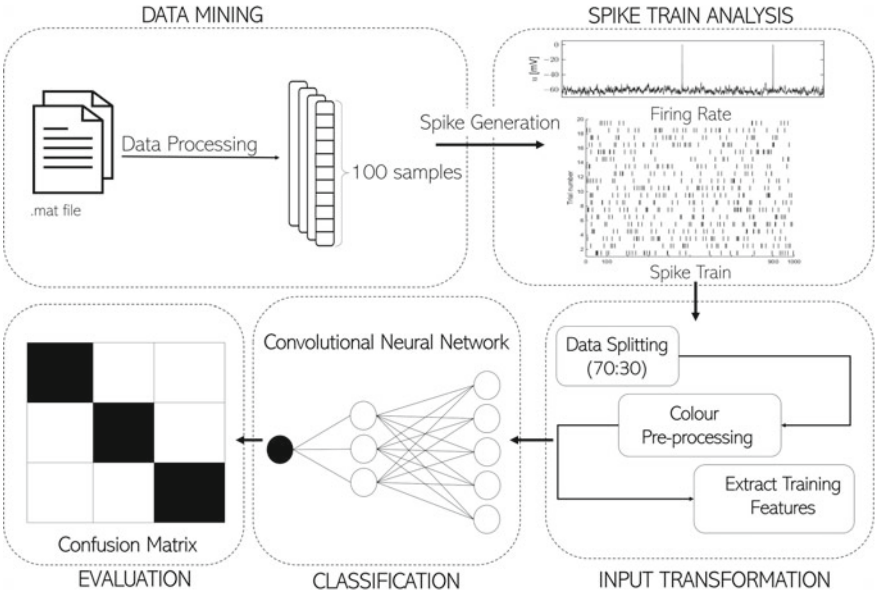


Fig. 2 General framework of proposed algorithm

4.2 Dataset

This study This paper employs UI-PRMD dataset [23], which consists 10 rehabilitation movements performed by 10 healthy individuals. Each person performed each movement repetitively 10 times in front of two sensory systems for motion capturing, which were Vicon optical tracker and a Kinect camera. The data is collected as positions and angles of the body joints in skeletal models provided by the sensors itself. However, this paper will be focused on Kinect camera angles data with 5 selected rehabilitation movements; Deep Squat, Hurdle Step, Inline Lunge, Sit to Stand, and Standing Shoulder Extension.

4.3 Spike Train Analysis

Data has been parted into 100 frames for each sample and encoded into firing rate and spike train. Figure 3 shows the comparison of the correctness for the movements that the firing spikes generated along 0° till 360° orientation direction. The incorrect movements might be assembled by incorrect angles of body joints. Hence, they did not achieve the target angle for perfect movements. From Fig. 4, each of the movements being recognized have their unique spike patterns, which reflect some aspects of neural functioning such as relating neural activity to stimuli, finding a repetitive pattern in a motor discharge and functional interactions between the neurons.

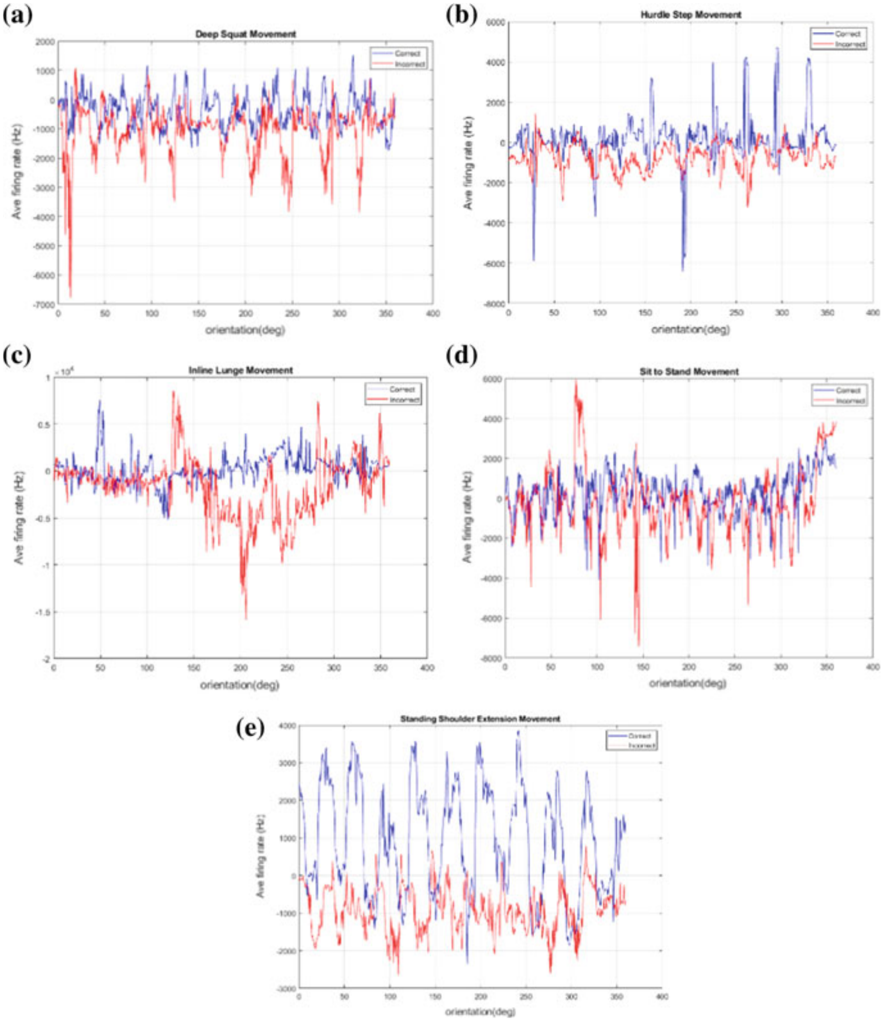


Fig. 3 Comparison of movements' correctness for: **a** Deep Squat; **b** Hurdle Step; **c** Inline Lunge; **d** Sit to Stand; **e** Standing Shoulder Extension

4.4 CNN Classification

Convolutional Neural Network is one of the most popular algorithms for deep learning, a type of machine learning in which a model learns to perform classification task directly from images, text, sounds, etc. In this paper, CNN has been adopted to classify spike trains pattern according to the type of the movements performed. Figure 5 illustrates the accuracy of the classification for the movements based on spike trains analysis classified by CNN. Sit to Stand exercises are most

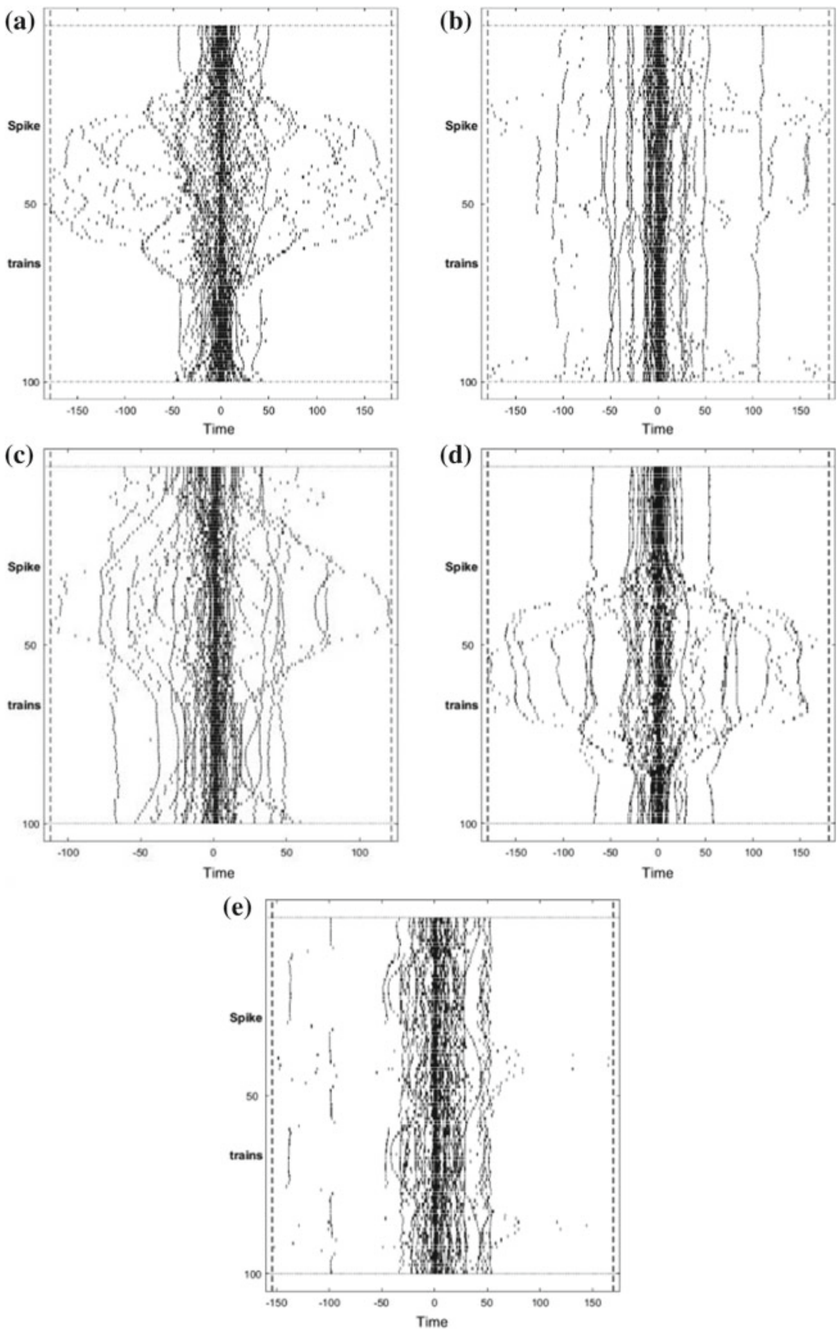
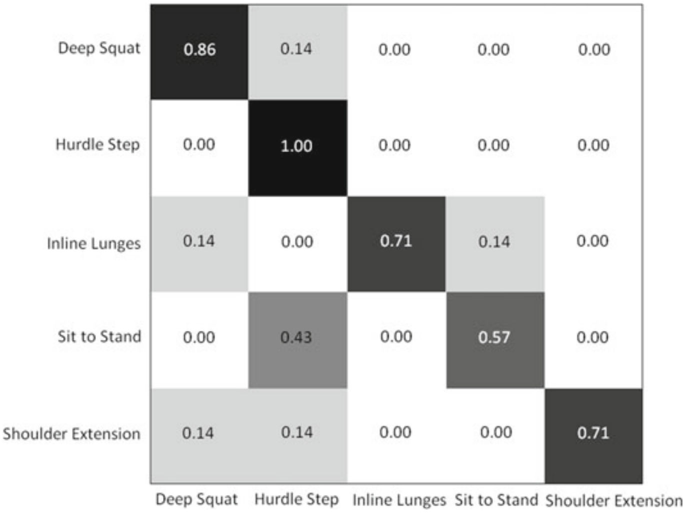


Fig. 4 Example of spike train of each movement: **a** Deep Squat; **b** Hurdle Step; **c** Inline Lunge; **d** Sit to Stand; **e** Standing Shoulder Extension



	Deep Squat	Hurdle Step	Inline Lunges	Sit to Stand	Shoulder Extension
Deep Squat	0.86	0.14	0.00	0.00	0.00
Hurdle Step	0.00	1.00	0.00	0.00	0.00
Inline Lunges	0.14	0.00	0.71	0.14	0.00
Sit to Stand	0.00	0.43	0.00	0.57	0.00
Shoulder Extension	0.14	0.14	0.00	0.00	0.71

Fig. 5 Confusion Matrix for classification accuracy of rehabilitation exercises movement

likely inverted with Hurdle Step with 43% similarity, as both of the exercises have identical lower limb movements as well as their spike train patterns. Other than that, Deep Squat and Shoulder Extension are both diverged by 14% each into Hurdle Step. However, Inline Lunges appear to be diverged 14% on Deep Squat and Sit to Stand each, respectively. Overall, the classification achieved 0.77 accuracy as the discrepancy between patterns, since all of them were mostly similar with Hurdle Step.

5 Conclusion

In this paper, we proposed spike train analysis for deep learning algorithm in rehabilitation exercise monitoring application. We achieved 77% accuracy of confusion matrix for classification part. By choosing spike trains, the features make a worthwhile contribution towards deep learning and they clearly can distinguish each of the exercises by their unique patterns. However, the algorithm needs to be revised to achieve the desired results. Hence, further filtering pre-processing will be implemented to strengthen the proposed method. This method also will be tested with other data to validate the performance to enhance the success rate of the accuracy. In future, we will compare with the other approaches in terms of the classification accuracy. As a conclusion, this proposed method works as we expected, however, it needs to be improved for further achievements.

Acknowledgements This work was supported by the Ministry of Higher Education (MoHE), Malaysia, under FRGS grant of Universiti Tun Hussein Onn Malaysia (No. 1584)

References

1. Rashid FN, Suriani NS, Nazari A (2018) Kinect-based physiotherapy and assessment: a comprehensive review. Indones J Electr Eng Comput Sci 11(3)
2. Krizhevsky A, Sutskever I, Hinton GE (2017) ImageNet classification with deep convolutional neural networks. Commun ACM 60(6):84–90
3. Kasabov N, Dhoble K, Nuntalid N, Indiveri G (2013) Dynamic evolving spiking neural networks for on-line spatio- and spectro-temporal pattern recognition. Neural Netw 41:188–201
4. Lucas P (2004) Bayesian analysis, pattern analysis, and data mining in health care. Curr Opin Crit Care 10(5):399–403
5. Zhu M, Chen W, Hirdes JP, Stolee P (2007) The K-nearest neighbor algorithm predicted rehabilitation potential better than current clinical assessment protocol. J Clin Epidemiol 60 (10):1015–1021
6. Zhu M, Zhang Z, Hirdes JP, Stolee P (2007) Using machine learning algorithms to guide rehabilitation planning for home care clients. BMC Med Inform Decis Mak 7:41
7. Muniz AMS et al (2010) Comparison among probabilistic neural network, support vector machine and logistic regression for evaluating the effect of subthalamic stimulation in Parkinson disease on ground reaction force during gait. J Biomech 43(4):720–726
8. Patsadu O, Nukoolkit C, Watanapa B (2012) Human gesture recognition using Kinect camera. In: 2012 ninth international conference on computer science and software engineering (JCSSE), pp 28–32
9. Suriani NS (2016) Fall detection using visual cortex bio-inspired model for home-based physiotherapy system BT—advances in machine learning and signal processing, pp 47–57
10. Ciresan DC, Meier U, Masci J, Gambardella LM, Schmidhuber J (2011) Flexible, high performance convolutional neural networks for image classification. In: Proceedings of the twenty-second international joint conference on artificial intelligence, vol 2, pp 1237–1242
11. Toshev A, Szegedy C (2014) DeepPose: human pose estimation via deep neural networks. In: Proceedings of the 2014 IEEE conference on computer vision and pattern recognition, pp 1653–1660
12. Du Y, Wang W, Wang L (2015) Hierarchical recurrent neural network for skeleton based action recognition. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR), pp 1110–1118
13. Gerstner W, Kistler WM (2002) Spiking neuron models: single neurons, populations, plasticity
14. Tavanaei A, Maida AS (2017) A spiking network that learns to extract spike signatures from speech signals. Neurocomputing 240:191–199
15. Bawane P, Gadariye S, Chaturvedi S, Khurshid AA (2018) Object and character recognition using spiking neural network. In: Proceeding materials today, vol 5, no 1, pp 360–366
16. Rongala UB, Mazzoni A, Oddo CM (2017) Neuromorphic artificial touch for categorization of naturalistic textures. IEEE Trans Neural Netw Learn Syst 28(4):819–829
17. Spigler G, Oddo CM, Carrozza MC (2012) Soft-neuromorphic artificial touch for applications in neuro-robotics. In: 2012 4th IEEE RAS & EMBS international conference on biomedical robotics and biomechatronics (BioRob), pp 1913–1918
18. Lecun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324
19. Szegedy C et al (1998) Visualizing and understanding convolutional networks. CoRR 86 (11):2278–2324

20. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. CoRR 1409(1)
21. He K, Zhang X, Ren S, Sun J (2015) Deep residual learning for image recognition. CoRR 1512(0)
22. Huang G, Liu Z, Weinberger KQ (2016) Densely connected convolutional networks. CoRR 1608(0)
23. Vakanski A, Jun H-P, Paul D, Baker R (2018) A data set of human body movements for physical rehabilitation exercises. Data 3(1):2

Faster Convergence of Q-Learning in Cognitive Radio-VANET Scenario

Mohammad Asif Hossain, Rafidah Md Noor, Saaidal Razalli Azzuhri, Muhammad Reza Z'aba, Ismail Ahmedy, Shaik Shabana Anjum, Wahidah Md Shah and Kok-Lim Alvin Yau

Abstract Cognitive Radio (CR) based Vehicular Ad hoc Network (VANET) or CR-VANET has become a very promising research domain. VANET is used to reduce road accidents, traffic congestion, and to provide other user experiences such as uninterrupted entertainment services. CR, on the other hand, solves bandwidth scarcity issue of VANET. For the high-speed mobility of the vehicles, the cognitive process of CR faces several challenges. Machine Learning (ML) has arrived as an integral tool to handle such challenges. Q-learning algorithm, a member of Reinforcement Learning (RL), which is a type of ML, is the most suitable for CR-VANET as it does not need any prior environment model and training dataset. But the problem is that it takes a longer time for learning purposes. In this paper, a dynamic ML framework is proposed. Case-based reasoning learning, cooperative spectrum sensing, teacher-student transfer learning approach will be aligned with the Q-learning for the faster convergence regarding the spectrum sensing issues in CR-VANET. The framework will accelerate the learning of the vehicles, and that is very important for the energy-efficient and real-life VANET implementation.

Keywords VANET · Cognitive radio · Reinforcement learning · Transfer learning · Q-learning · Case-based reasoning

M. A. Hossain · R. M. Noor (✉) · S. R. Azzuhri · M. R. Z'aba · I. Ahmedy · S. S. Anjum
Faculty of Computer Science and Information Technology, University of Malaya,
Kuala Lumpur, Malaysia
e-mail: fidah@um.edu.my

R. M. Noor
Centre for Mobile Cloud Computing Research, University of Malaya,
Kuala Lumpur, Malaysia

W. M. Shah
Faculty of Information and Communication Technology, Universiti Teknikal Malaysia
Melaka, Melaka, Malaysia

K.-L.A. Yau
School of Science and Technology, Sunway University, Selangor, Malaysia

Z. Zakaria and R. Ahmad (eds.), *Advances in Electronics Engineering*,
Lecture Notes in Electrical Engineering 619,
https://doi.org/10.1007/978-981-15-1289-6_16

1 Introduction

Every year, around 1.25 million people die from the road accidents [1] and the resulting congestion incurs a huge amount of money (In the U.S. alone, congestion cost $305 billion in 2017 [2]). VANET has emerged as a solution to these alarming situations. For implementing VANET, a gigantic amount of real-time data (such as of GPS (Global Positioning System), radar, camera, LIDAR (Light Detection and Ranging), Sonar, sensors data), and infotainment data will be exchanged in the coming years. According to Intel, each smart vehicle is going to generate and consume approximately 4 terabytes of data in on average per day driving by 2020 [3].

IEEE 802.11p or IEEE 1609, also known as Dedicated Short-Range Communications (DSRC) standard is reserved for the vehicular networks with 75 MHz bandwidth in the frequency range of 5.85 to 5.925 GHz. This allocated bandwidth is not sufficient enough to accommodate such a massive amount of data [4]. On the other hand, licensed bandwidths or frequencies such as TV band or military radio band are not properly utilized [5]. The report shows that more than 60% bandwidth of below 6 GHz spectrum is not being used or not properly utilized [6]. CR, the concept coined by Mitola & Maguire [7], has emerged as the solution in the bandwidth scarcity problem. CR users are allowed to sense and use these underutilized licensed channels dynamically in an opportunistic manner, as well as for spectrum mobility that allows users to vacate licensed channels re-occupied by licensed users (primary users or PUs). The latency for the safety message exchange must be lower than 100 ms, but in general, the cognitive processes takes around 2 s time [8]. Moreover, a huge amount of network overhead is transmitted due to such cognitive processes and unnecessary repetitive computational tasks have to be performed. These will lead to unnecessary energy consumption.

ML can be applied in CR-VANET to make it more intelligent to adapt the uncertain radio environment to solve those issues. It can ensure faster decision, reliability, energy efficiency, and enhanced QoS [9]. There are three main categories of ML techniques: supervised learning, unsupervised learning, and RL. Other learning methods such as semi-supervised learning, online learning, and transfer learning are the variation of these three categories [10]. Q-learning, a type of among several RL algorithms, is found as the most suitable for the CR-VANET scenario due to its adaptability with the dynamic environment, model-free requirement, and working capability without training dataset [11]. Here, agents face an unpredictable environment by selecting appropriate actions by using mathematical approaches and receiving rewards consequently. The main issue faced by a Q-learning agent is that it takes longer learning phases, i.e. a huge number of iterations are required for the convergence. This is due to its learning itself all alone. In this paper, a dynamic learning framework has been proposed. The objective of this framework is to reduce the overall learning time of the vehicles about the vacant spectrums on the surrounded environment. The idea of teacher-student approach (a type of transfer learning which is a feature of ML) along with case-based reasoning (CBR) will accelerate the learning time of Q-learning. In a teacher-student approach, an already learned vehicle (teacher) will share its own sensing information to the learning vehicle (student) [12]. CBR, another type of ML, tries to solve new problems by reusing past solutions that were used to solve similar problems. This cognitive process uses prior stored 'case' (results and experience) to fit a new similar problem situation [8].

The remainder of this paper is organized as follows: Sect. 2 discusses the related works, Sect. 3 provides the overview and the problem formulation of Q-learning, Sect. 4 discusses the proposed framework, Sect. 5 describes the performance evaluation methods and parameters, and finally, Sect. 6 concludes the paper.

2 Related Works

Several works were done in the fields of spectrum sensing in CR-VANETs by using Q-learning. In [13], the author proposed architecture by using Q-learning and CBR for VANET to enable automatic learning of the radio environment by the vehicles. The authors in [14] showed that by using this learning, the total energy consumption due to the spectrum sensing can be reduced to only about 1.72% compared to the traditional spectrum sensing method. In [15], the authors used deep Q-learning for designing an optimal data transmission scheduling scheme in CR-VANET to minimize transmission costs. They used cache memory for taking the decision. Their scheme's convergence took place after 13,000 to 20,000 iterations at 28 m/s vehicle speed. Morozs et al. in [16] proposed a scheme, which integrates distributed Q-learning and CBR aimed to facilitate a number of learning processes running in parallel. They got the best result after 1,000,000 iterations. RL method was considered for the CR network with RF energy harvesting in [17]. Their proposed scheme was for the optimum switching between the transmit mode, energy harvesting mode, and receiving mode of the CR users. They got average throughput converges to 0.68 after 1,000,000 iterations.

The above-mentioned works found very good performance in spectrum sensing in terms of higher probability of PUs detection with a lower probability of false alarm, but with a very slow convergence rate. They needed a huge number of iteration to learn the environment optimally. For the practical point of view, these learning times are quite infeasible. Authors in [18] gave some insights about the way to accelerate Q-learning time, though it was theoretical and was not considered the aspects of CR-VANET. This paper is targeted to reduce such learning time (i.e. make the convergence faster).

3 Q-Learning Algorithm

Q-learning, the most used type of RL, is an on-line algorithm, which enables an agent to learn in an interactive manner with its surrounding environment. The main aim of Q-learning is to exploit the long-term rewards receiving in the future. It does not require any environment model and dataset for the training. In Q-learning, an agent or the learner (say a CR based vehicle) is interacting with the radio environment (comprising everything outside the agent).

From Fig. 1, it is shown that, at each step t, the agent observes the state of its surrounding environment $s_t \epsilon S$, where S is a set of possible states. Based on knowledge gained at s_t, the agent selects an action $a_t \epsilon A$, where A is a set of actions. At the next step $t + 1$, the environment transits to a new state s_{t+1} and the agent gets a reward of r_t. Based on the reward table, the agent chooses the next action (it may be beneficial or may be harmful) and then they update a new value called Q-value mapping of state-action pairs Q (s_t, a_t). Several Q-values are stored in the Q-table. For example, in CR-VANET scenario, an action might be choosing any spectrum for accessing, the state might be the location and time of the vehicle. If the sensed spectrum faces interferences by the PUs, the agent would get a negative reward, otherwise gets a positive reward.

After every action, the agent gets the reward and updates its Q-value based on Eq. (1).

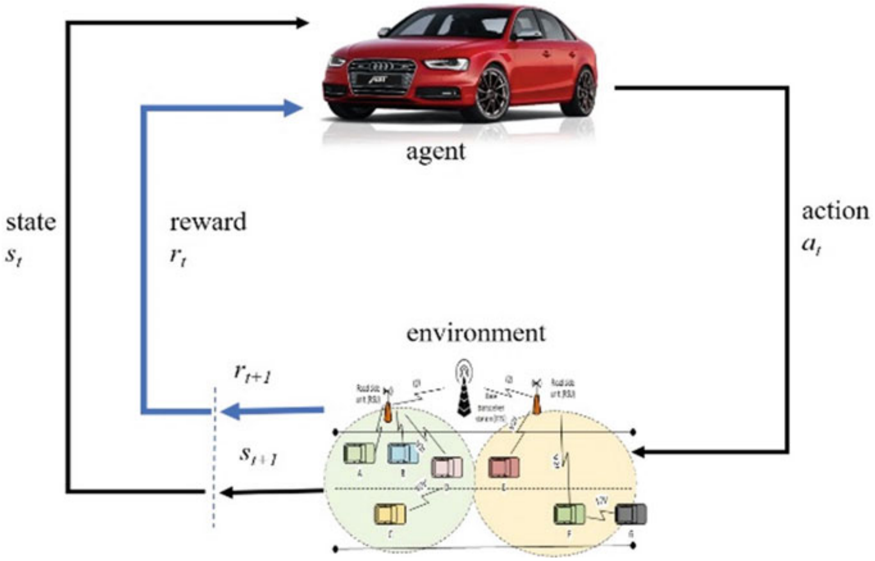


Fig. 1 Q-learning approach

$$\begin{aligned} Q_{new}(state, action) \leftarrow & (1-\alpha) Q_{old}(state, action) \\ & + \alpha(reward + \gamma \, max \, Q_{old}(next\,state, all\,actions)) \end{aligned}$$

$$\because \quad Q_{t+1}(s_t, a_t) \leftarrow (1-\alpha)Q_t(s_t, a_t) + \alpha\left[r_{t+1}(s_{t+1}, a_t) + \gamma \max_{a \in A} Q_t(s_{t+1}, a)\right] \tag{1}$$

Here,

α: The *learning rate*, which determines how much the new Q-value overrides the previous Q-value. α ranges from 0 to 1. The higher value of α means the higher speed of the learning process (may lead to faster convergence), but sometimes stability is lost and failed to converge. The lower the value of α, smoother the learning process but slower rate of convergence.
γ: The *discount factor*, which implies how much importance is given to future rewards.
r: The *reward* received by the agent. The short-term reward is called the *delayed reward* and the future reward is called the *discounted reward.*

There are two policies for taking action. When the agent chooses for exploitation (uses existing knowledge to select the best action), it uses an optimal policy and when it chooses for exploration (needs more knowledge), it uses a random policy. The agent receives positive delayed rewards when it selects a proper action for a particular state. Positive value increases and the respective Q-value, and vice versa [19]. Therefore, the target of Q-learning is to get an optimal policy (agent behavior) π*:* $S \rightarrow A$, which can maximize the reward at state S [20].

The optimal Q-value for a particular state can be written as:

$$V^{\pi^*}(s_t) = \max_{a \in A} Q_t(s_t, a) \tag{2}$$

Therefore, the optimal policy can be written as:

$$\pi^*(s_t) = \arg\max_{a \in A} Q_t(s_t, a) \tag{3}$$

From the above discussions, it is clear that the convergence rate depends on the quality of Q-table and the value of α and γ. The more reward an agent accumulated, the better Q-table would get, and therefore, the convergence will be faster. But the issue is Q-learning algorithm is learning totally by itself, not taking any helps from others. For better performance and convergence, it has to face the tradeoff between exploration and exploitation. More exploration provides better decision (sacrifices immediate rewards hoping for more future rewards), but slower convergence, on the other hand, quick exploitation might provide faster convergence, but poorer performance. If the Q-table is updated with more rewarded state-action pairs, overall convergence would be faster.

4 Proposed Dynamic Machine Learning Framework

In this paper, a dynamic ML framework that includes Q-learning and CBR has been developed. Teacher-student transfer learning approach has also been used in the framework. Here, a learned vehicle (teacher) shares its own sensing information to the learning vehicle (student) [12]. The vehicle chooses the best ML based on the proposed framework. Suppose, if the user chooses the same known path at the same time of the day, the CBR would be used, and if the environment is unknown to the CBR-database, Q-learning method would be used.

In this proposed theme, like teacher-student approach in [12], a learned vehicle, for example, might have the best action-state pair or best Q(s_t, a_t). If the learning vehicle is getting Q value from this learned vehicle, it does not need additional exploration for that state. For example, in Fig. 2, a learning car (A) has broadcasted a request for the spectrum sensing information to the neighbor vehicles. A is in say s_{tk} state, on the request it will include this state value. A teacher (say B) has the best-rewarded information regarding state s_{tk}, so it will then forward Q(s_{tk}, a_{tk}) as the response to A. Another car (C) say, for example, does not have information regarding the s_{tk} state. So, it would not respond. After getting the Q(s_{tk}, a_{tk}) from B, the car A will keep this Q value to its Q-table. So, in future, when car A is in the same state (s_{tk}), it would not go for exploration state. In this way, by cooperation, a learning vehicle can increase its learning process. Figure 3 shows the proposed framework. This framework will provide faster convergence and reduce the overall sensing time, hence, provides the energy efficient and improved QoS CR-VANET. This framework is also described in Algorithm 1 and the Q-learning algorithm in Algorithm 2. Here, when a vehicle selects the destination and starts its journey, it will search its own database whether the route (the road) is already known or unknown. The i_{th} vacant spectrum information contains the location (l_i), time (t_i), and channel (c_i). If the information is found known by the searching database, it retrieves spectrum information from the database (learned previously) and uses that vacant channel.

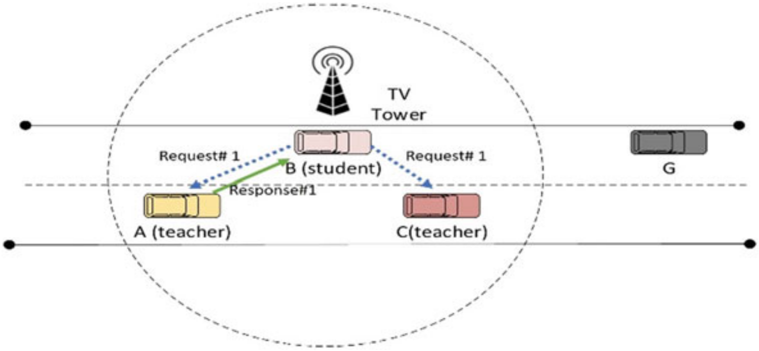


Fig. 2 Teacher-student transfer learning approach

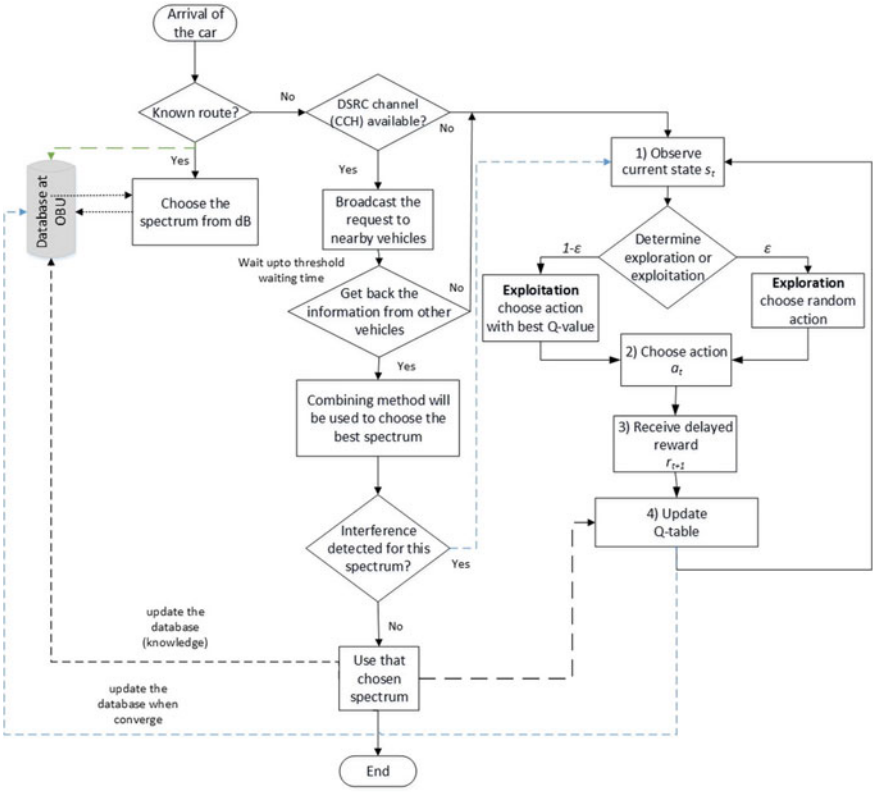
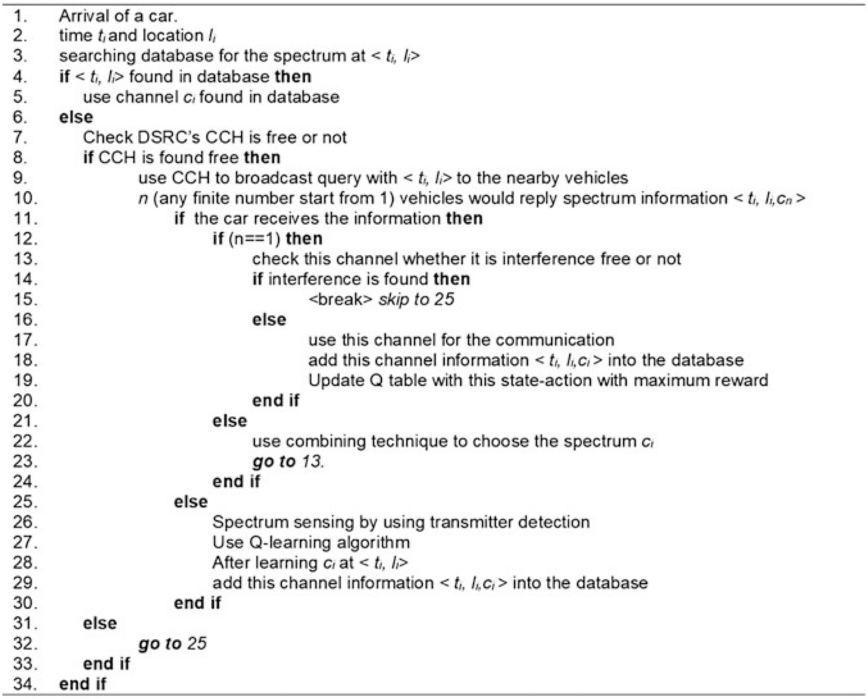

Arrival of the car
Known route?
No
DSRC channel (CCH) available?
No
Yes
Database at OBU
Choose the spectrum from dB
Yes
Broadcast the request to nearby vehicles
Wait upto threshold waiting time
Get back the information from other vehicles
No
Yes
Combining method will be used to choose the best spectrum
Interference detected for this spectrum?
Yes
No
update the database (knowledge)
Use that chosen spectrum
update the database when converge
End
1) Observe current state s_t
Determine exploration or exploitation
1-ε
ε
Exploitation choose action with best Q-value
Exploration choose random action
2) Choose action a_t
3) Receive delayed reward r_{t+1}
4) Update Q-table

Fig. 3 Proposed framework of dynamic machine learning in CR-VANET

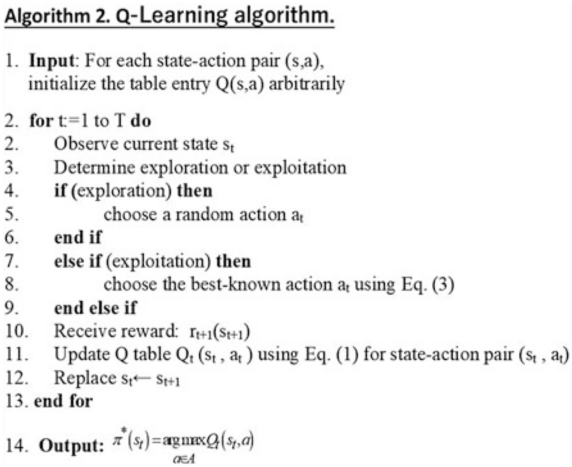
If the route is found unknown, the vehicle will look for whether the DSRC's common control channel (CCH) is available or not. If it's found available, it broadcasts a query message to its neighbor vehicles for learning the free channel of that route on that time. If the vehicle gets the responses from several vehicles, it will use any suitable combining method (like maximal-ratio combining or MRC) to choose the best channel. It will then test whether the channel is really free or not by using any detection method. If it finds interference-free, it will use that channel and stores this information (l_i, t_i, c_i) to the database and to the Q-table. If the vehicle finds CCH unavailable or detects interference or does not get information from any vehicle, it will go for non-cooperative spectrum sensing by using a primary transmitter detection method. Q-learning will be used for taking further action and get backs as rewards/punishments. Here, the action means selecting the spectrum to use, the agent gets a reward when it finds interference-free (absence of PUs) and gets punishment when it finds interference on its chosen spectrum. After some iterations, it will be converged and then updates the database. It will add (l_i, t_i, c_i) into the database. For the Q-learning, the ε-greedy policy has been considered.

In, ε-greedy policy, the agent chooses exploration with a small probability ε and exploitation with probability (1 − ε).

Algorithm 1 Dynamic Spectrum Sensing



```
1.  Arrival of a car.
2.  time $t_i$ and location $l_i$
3.  searching database for the spectrum at < $t_i$, $l_i$>
4.  if < $t_i$, $l_i$> found in database then
5.      use channel $c_i$ found in database
6.  else
7.      Check DSRC's CCH is free or not
8.      if CCH is found free then
9.              use CCH to broadcast query with < $t_i$, $l_i$> to the nearby vehicles
10.             $n$ (any finite number start from 1) vehicles would reply spectrum information < $t_i$, $l_i$,$c_n$ >
11.                 if the car receives the information then
12.                     if (n==1) then
13.                         check this channel whether it is interference free or not
14.                         if interference is found then
15.                                 <break> skip to 25
16.                         else
17.                                 use this channel for the communication
18.                                 add this channel information < $t_i$, $l_i$,$c_i$ > into the database
19.                                 Update Q table with this state-action with maximum reward
20.                         end if
21.                     else
22.                         use combining technique to choose the spectrum $c_i$
23.                         go to 13.
24.                     end if
25.                 else
26.                     Spectrum sensing by using transmitter detection
27.                     Use Q-learning algorithm
28.                     After learning $c_i$ at < $t_i$, $l_i$>
29.                     add this channel information < $t_i$, $l_i$,$c_i$ > into the database
30.                 end if
31.     else
32.             go to 25
33.     end if
34. end if
```



Algorithm 2. Q-Learning algorithm.

```
Input: For each state-action pair (s,a),
initialize the table entry Q(s,a) arbitrarily

for t:=1 to T do
    Observe current state $s_t$
    Determine exploration or exploitation
    if (exploration) then
            choose a random action $a_t$
    end if
    else if (exploitation) then
            choose the best-known action $a_t$ using Eq. (3)
    end else if
    Receive reward: $r_{t+1}(s_{t+1})$
    Update Q table $Q_t$ ($s_t$ , $a_t$ ) using Eq. (1) for state-action pair ($s_t$ , $a_t$)
    Replace $s_t \leftarrow s_{t+1}$
end for

Output: $\pi^*(s_t) = \underset{a \in A}{\arg\max} Q_t(s_t, a)$
```

The overall learning will not be solely depending on previous data, not on CSS only nor on Q-learning only. The proposed learning is a kind of hybrid learning and will eventually become faster and more reliable. The main concept here is that, at first step, every vehicle searches its own database, if it finds, it uses that vacant frequency. If it does not find information from its own database, it would ask for help from other surrounding vehicles. The vehicles that already know about the sensing information (on that route and on that time) will deliver their learned sensing information to that vehicle. Here, the teacher-student transfer learning approach has been used. After getting the information from several vehicles, the vehicle would use fusion or combining technique. Learned sensing information will be fed to the database and to the Q-table. If all these stages fail, it will go for the non-cooperative SS and RL phases. This system provides faster spectrum information and increases the convergence rate.

5 Performance Evaluation

For the performance evaluation purposes, SUMO (Simulation of Urban MObility) simulator, Network simulator 3 (NS3), and Python programming will be used. Figure 4 shows the steps of getting the results for the performance evaluation. By using SUMO, the real-life mobility model and VANET will be designed, then this will be integrated with the NS3 to add the feature of CR. After running the simulation, spectrum data of CR-VANET would be obtained. These data will be fed to the Python, wherein the framework of Q-learning, CBR, teacher-student, and CSS would be implemented. After performing data analysis by Python, the results would be obtained.

Following performance metrics would be used for the performance measurements of the proposed framework:

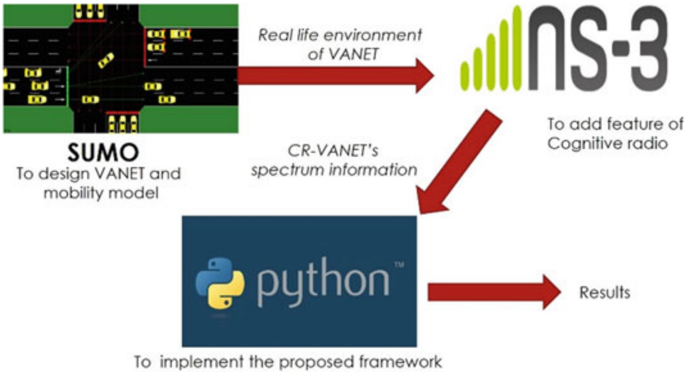

Real life environment of VANET
NS-3
To add feature of Cognitive radio
SUMO
To design VANET and mobility model
CR-VANET's spectrum information
python
Results
To implement the proposed framework

Fig. 4 Steps in performance evaluation

Convergence rate: it defines how fast an agent (vehicle) learns the surrounded complex environment or simply the number of iterations needed for an algorithm to start providing the best optimal value. The faster the convergence rate (less iteration required), the better the algorithm performs. The aim of this paper is to make the convergence faster (learns the system within a short period of time).

The probability of false alarm versus the probability of detection: the probability of false alarm means the probability of declaring about the presence of a PU, though that sensed spectrum is not really occupied by any PU. On the other hand, the probability of detection represents the probability declaring the presence of a PU and that sensed spectrum is truly occupied by that PU. In CR-VANET, it is one of the most used performance metrics.

Energy efficiency: It is the measurement by which the performance of a system can be evaluated. When a system provides the same services but with less energy consumption compared to other systems, then it can be said that the earlier system performs better in terms of energy and it is an energy efficient system.

Delay: It is one of the major issues in CR-VANET scenario. It is the difference between the theoretical time taken by a system and the actual time it takes to perform any task. The cognitive process should be performed with a very lower delay. High delay reduces the overall performance of a system.

This research work is expecting a higher probability of detection and lower probability of false alarm, faster convergence, higher energy efficiency, and lesser delay compared to the existing techniques and methods for the spectrum sensing in CR-VANET scenario.

6 Conclusion

Vehicular Ad hoc Network (VANET) has emerged as one of the major solutions to enhance road safety, reduce traffic congestion, and improve quality-of-service (QoS). Cognitive Radio (CR), on the hand, has appeared to alleviate the spectrum scarcity issue of exponentially growing VANETs. Machine learning tools are now becoming an integral part of CR-VANET to boost its advantages. In this paper, a dynamic machine learning framework has been proposed. The framework consists Q-learning, case-based reasoning, and teacher-student transfer learning concept. The proposed framework is expected the improvement in terms of convergence rate. The proposed method is also expected to provide reliable learning to the vehicles in very dynamic environments with reduced delay and network overhead. In future work, we will analyze, design, and validate this proposed dynamic machine learning framework with considering various challenges such as PU activity models, hidden PUs problem, Doppler effects and so on.

Acknowledgements This work was supported by the Partnership Grant under Grant RK004-2017 and Grant CR-UM-SST-DCIS-2018-01 between University of Malaya and Sunway University.

References

1. WHO (2015) Road safety report 2015 (Online). Available: http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/
2. Schneider B (2018) Traffic's mind-boggling economic toll (Online). Available: https://www.citylab.com/transportation/2018/02/traffics-mind-boggling-economic-toll/552488/
3. Winter K (2017) For self-driving cars, there's big meaning behind one big number: 4 terabytes (Online). Available: https://newsroom.intel.com/editorials/self-driving-cars-big-meaning-behind-one-number-4-terabytes/
4. Singh KD, Rawat P, Bonnin J-M (2014) Cognitive radio for vehicular ad hoc networks (CR-VANETs): approaches and challenges. EURASIP J Wirel Commun Netw 2014(1):49
5. Vo QD, Choi JP, Chang HM, Lee WC (2010) Green perspective cognitive radio-based M2M communications for smart meters. In: 2010 international conference on information and communication technology convergence, ICTC 2010, pp 382–383
6. Spectrum Policy Task Force Report: FCC, in Cambridge University Press, 2012, 2002
7. Mitola G, Maguire J (1999) Cognitive radio: making software radios more personal. IEEE Pers Commun 6(4):13–18
8. Chen S, Vuyyurut R, Altintas O, Wyglinski AM (2011) Learning in vehicular dynamic spectrum access networks : opportunities and challenges. In: 2011 International Symposium on Intelligent Signal Processing and Communications Systems, pp 5–10
9. Yau KLA, Komisarczuk P, Teal PD (2010) Applications of reinforcement learning to cognitive radio networks. In: 2010 IEEE International Conference on Communications Workshop, pp 1–6
10. Liang L, Ye H, Li GY (2018) Towards intelligent vehicular networks: a machine learning framework. IEEE Internet Things J 6(1)
11. Wu C, Ohzahata S, Kato T (2013) Flexible, portable, and practicable solution for routing in VANETs: a fuzzy constraint Q-Learning approach. IEEE Trans Veh Technol 62(9):4251–4263
12. Da Silva FL, Glatt R, Costa AHR (2017) Simultaneously learning and advising in multiagent reinforcement learning. In: Proceedings of the 16th conference on autonomous agents and multiagent systems, pp 1100–1108
13. Chen S, Vuyyuru R, Altintas O, Wyglinski AM (2011) On optimizing vehicular dynamic spectrum access networks: automation and learning in mobile wireless environments. In: IEEE vehicular networking conference, VNC, pp 39–46
14. Grace D, Chen J, Jiang T, Mitchell PD (2009) Using cognitive radio to deliver 'green' communications. In: Proceedings 2009 4th international conference on cognitive radio oriented wireless networks and communications, CROWNCOM 2009, pp 2–7
15. Zhang K, Leng S, Peng X, Pan L, Maharjan S, Zhang Y (2018) Artificial intelligence inspired transmission scheduling in cognitive vehicular communications and networks. IEEE Internet Things J 6(2)
16. Morozs N, Clarke T, Grace D (2016) Cognitive spectrum management in dynamic cellular environments: a case-based Q-learning approach. Eng Appl Artif Intell 55:239–249
17. Van Huynh N, Hoang DT, Nguyen DN, Dutkiewicz E, Niyato D, Wang P (2018) Reinforcement learning approach for RF-powered cognitive radio network with ambient backscatter. In: CoRR, vol abs/1808.0, pp 1–6
18. Potapov A, Ali MK (2003) Convergence of reinforcement learning algorithms and acceleration of learning. Phys Rev E 67(2):26706
19. Ling MH, Yau K-LA, Qadir J, Poh GS, Ni Q (2015) Application of reinforcement learning for security enhancement in cognitive radio networks. Appl Soft Comput J 37:809–829
20. Sutton RS, Barto AG (2017) Reinforcement learning: an introduction, 2nd edn. The MIT Press

Using Multicultural Herbal Information to Create Multi-pattern Herb Name Retrieval System

Verayuth Lertnattee and Narinee Phosri

Abstract In Thailand, many botanical gardens are attracted by local and foreign visitors. Searching herbal information in these gardens usually applies plant names as keywords. However, various patterns of names can be used, e.g., common names, local names for each language, and scientific names. Moreover, misspelling and different spelling terms are usually found. Information of plants in gardens is often given in Thai and/or English. Searching information by inputting Thai plant names may be unfamiliar to foreign visitors. To help visitors retrieve herbal information by herb names, a multi-pattern herb name retrieval system was implemented. Using a multicultural herbal information system (HerbINFO), an approximate string matching technique, and a machine translator, visitors can easily search herbal information by any patterns of herb names. Using herb names contributed from members of the HerbINFO, feedbacks from a mobile phone and a voting system in the HerbINFO, efficiency of searching will be dynamically improved. A set of herb names in several patterns and languages was used to demonstrate the system via a mobile application.

Keywords Multiculture · Herb name · String matching technique · Levenshtein · Machine translator

1 Introduction

Thailand is the country having plenty of botanical gardens to visit as tourist attractions. Those gardens regularly inform the plant names in their database as the Thai language. Searching plant information commonly uses a plant name as a key word [1]. However, a plant name generally has various types e.g., common names,

V. Lertnattee (✉)
Faculty of Pharmacy, Silpakorn University, Nakhon Pathom 73000, Thailand
e-mail: lertnattee_v@su.ac.th

N. Phosri
Sirindhon College of Public Health, Suphanburi, Thailand

Z. Zakaria and R. Ahmad (eds.), *Advances in Electronics Engineering*,
Lecture Notes in Electrical Engineering 619,
https://doi.org/10.1007/978-981-15-1289-6_17

local names, and scientific names. The problem of searching plant data is the plant name can be written in various alphabets. In addition, users possibly enter the wrong characters because of the following reasons: misspelling, using different spelling terms, and using different tone marks [2]. Additionally, foreign visitors can be unfamiliar or inconvenient to use the Thai language to search information on the Thai plant database. To alleviate these problems, a multi-pattern herb name retrieval system was constructed based on the HerbINFO, a multicultural herbal information system which herb names are gathered from members of the system. However, herb names (common names and local names) have been often given in Thai. To improve efficiency of retrieving information by herb names, an approximate string matching technique and machine language translators are also applied. Information about the correct herb should be found even misspelling or different spelling keywords in any patterns of herb names. A set of herb names was used for demonstration by a frontend mobile application. Moreover, herb names and their languages may be contributed into the database to improve efficiency of retrieving in the future. In the rest of this paper, the background about herb names in multicultural herbal information, string matching techniques, and machine translators are given in Sect. 2. The detail of our proposed method is made in Sect. 3. Section 4 describes experimental settings. The experimental results are reported in Sect. 5. Section 6 provides conclusion and future work.

2 Background

2.1 Herb Names in Multicultural Herbal Information

Various names are used to represent an herb. A plant is usually found in a region, it should have its common names for languages used in that region. Besides a set of common names, local names for each language are also given by native people of particular area [3]. In Thailand, two alphabets are used in common, i.e., Thai and English. Therefore, several herbs, which are not originate in Thailand, their names are transliterated/transcribed from other languages to Thai and/or English. For example, names of Chinese herbs and crude drugs are transliterated/transcribed from Chinese to Thai and/or English. When we use a common name or a local name to present an herb, problems of synonyms and homographs may be occurred [4]. In this context, synonyms mean an herb can be presented by several names while homographs mean a name may refer to several herbs. When a common name/local name of an herb is used for searching, a list of herbs with the same name may be found. Users should select the correct herb they would like to find. To confirm the correct herb, the scientific name of an herb is used. The HerbINFO is created to collected herb names. Thai and English common names, including scientific names of herbs are given by administrators. Common names of other languages including local names for each language will be given by members of the HerbINFO. A voting system usually improves accuracy of the opinions [5]. With this mechanism, the correctness of herb names provided by members should be improved.

2.2 String Matching Algorithm

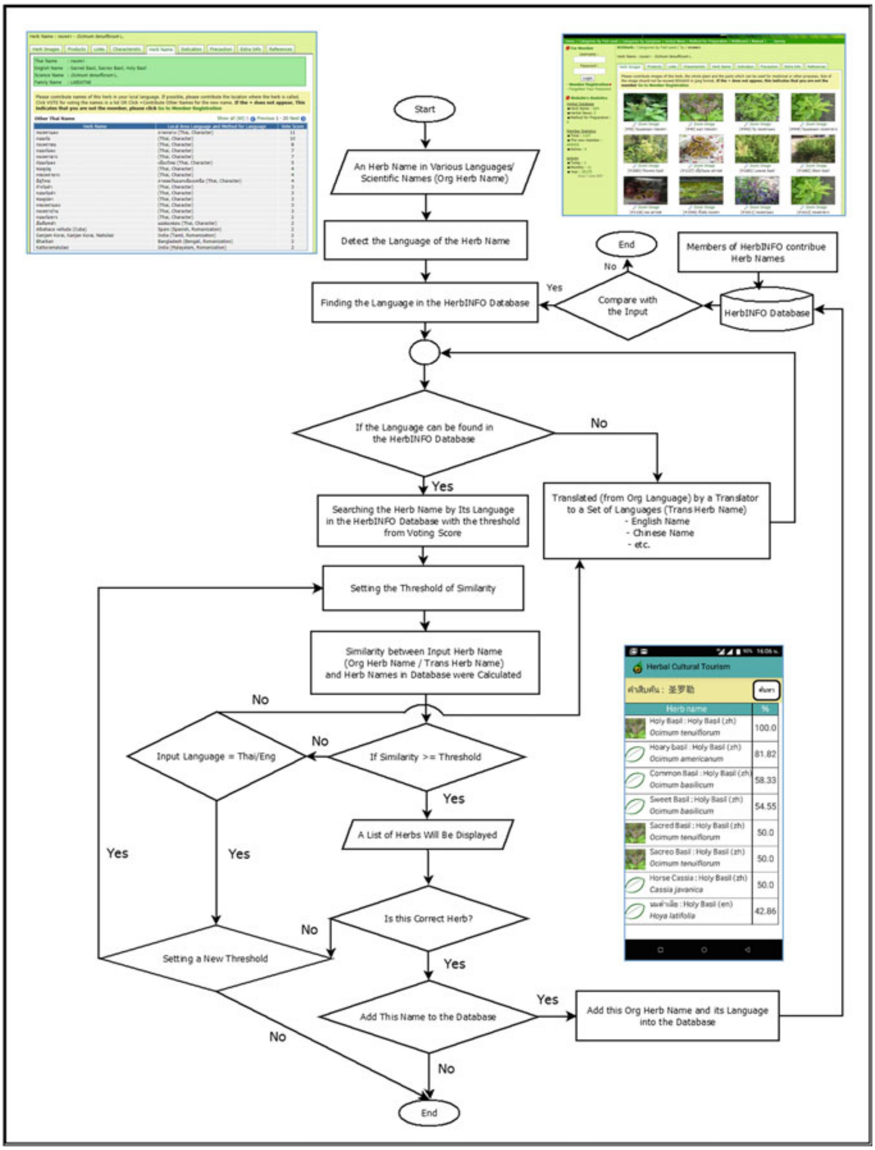
Besides the complexity of herb names as mention earlier, misspelling and different spelling words may usually found. An approximate string matching technique should be applied to alleviate problems of misspelling and different spelling inputs [6]. A well-known algorithm, i.e., Levenshtein algorithm [7], is useful for this purpose. Similarity between two strings by edit distance is calculated. Minimum value to transform string S to string T by insertion, deletion or replacement characters is indicated that S is similar to T. This function can be represented by Levenshtein (S, T, insertion, replacement, deletion), for short Levenshtein (i, r, d). The function to transform the edit distance to similarity (%Simlev) is (1-distance (S, T)/max (|S|, |T|)) *100. The distance(S, T) is the minimum edit distance for transforming S to T and max (|S|, |T|) is the maximum length of string between S and T. The similarity threshold level for similarity is set to filter only a set of potential strings we would like to retrieve. In this paper, S is an input herb name and T is an herb name from a collection of herb names stored in the HerbINFO database. With the voting system in HerbINFO, only high voting score herb names greater than or equal to the voting threshold (set by administrator) are used to compute similarity. The list of herbs, which is gained values of the similarity greater than or equal to the similarity threshold, is displayed.

2.3 Machine Translator

Machine translators are program used for translating texts or speech from one language into another language. They were developed to solve the problem of communication in different languages. They are available online, such as Google Translate, Bing Microsoft Translator, and Yandex Translate. Moreover, several of these programs have Application Programming Interface (API) which integrated the program with the translation service [8]. From statistics in 2017 from Ministry of Tourism and Sports [9], the top ranks of foreign visitors were shown. The popular languages used by foreign visitors were English, Chinese, Japanese, etc.

3 Multi-pattern Herb Name Retrieval System

In this system, herb names in the HerbINFO database is used for computing similarity with input herb names. Thai common names, English common names, and scientific names of herbs in the HerbINFO are given by administrators. Various


Start
An Herb Name in Various Languages/ Scientific Names (Org Herb Name)
Detect the Language of the Herb Name
Finding the Language in the HerbINFO Database
Yes
Compare with the Input
No
End
Members of HerbINFO contribue Herb Names
HerbINFO Database
If the Language can be found in the HerbINFO Database
No
Yes
Searching the Herb Name by Its Language in the HerbINFO Database with the threshold from Voting Score
Translated (from Org Language) by a Translator to a Set of Languages (Trans Herb Name) - English Name - Chinese Name - etc.
Setting the Threshold of Similarity
Similarity between Input Herb Name (Org Herb Name / Trans Herb Name) and Herb Names in Database were Calculated
No
No
Input Language = Thai/Eng
If Similarity >= Threshold
Yes
A List of Herbs Will Be Displayed
Yes
Yes
Is this Correct Herb?
No
Setting a New Threshold
Yes
Add This Name to the Database
Yes
Add this Org Herb Name and its Language into the Database
No
No
End
Herbal Cultural Tourism
คำสืบค้น : 圣罗勒
ค้นหา

Herb name	%
Holy Basil : Holy Basil (zh) *Ocimum tenuiflorum*	100.0
Hoary basil : Holy Basil (zh) *Ocimum americanum*	81.82
Common Basil : Holy Basil (zh) *Ocimum basilicum*	58.33
Sweet Basil : Holy Basil (zh) *Ocimum basilicum*	54.55
Sacred Basil : Holy Basil (zh) *Ocimum tenuiflorum*	50.0
Sacreo Basil : Holy Basil (zh) *Ocimum tenuiflorum*	50.0
Horse Cassia : Holy Basil (zh) *Cassia javanica*	50.0
นมตำเลีย : Holy Basil (en) *Hoya latifolia*	42.86

Fig. 1 Algorithm for searching an input herb name

local names in Thai are contributed from the HerbINFO's members. Some contributed names are English and some herb names are transliterated/transcribed from other languages, especially Chinese, to Thai/English. When an herb name is input into the system, the language of the input will be detected. Initial with the query for retrieving herb names, which its language is corresponding to the input language and the voting score greater than or equal to the voting threshold, the similarity between input herb names and herb names from query result, are calculated. If the similarity score is greater than or equal to the similarity threshold, a list of herbs will be displayed. Computing similarity is reprocessed if the language of the input herb name is not found in the database or a list of herbs does not display when the input language is other than Thai/English. Then, a machine translator is used to translate an original herb name (Org Herb Name) with input language to a translated herb name (Trans Herb Name) in a set of common languages, e.g., English, Chinese, etc. When a set of herbs is listed and users/visitors find the correct herb that they intend to search by compare the original input herb name to the scientific name and/or images of that herb, they can select the correct herb. The original input herb name may be considered to collect into the HerbINFO database. The value of similarity threshold can be set to limit or extend the list of herbs. Algorithm for searching an input herb name is shown in Fig. 1.

4 Experimental Settings

Two experiments were done, (1) herb names in Thai/English (including other languages that detect as English) and (2) herb names in other languages. For the first experiment, inputs were varied to Thai and English. Misspelling and different spelling terms were presented. In the second experiment, herb names with characters of four popular languages used by foreign visitors were input. All herb names were types into the mobile application. A language of an input was detected and translated to English by Yandex Translator [10]. The original and translated inputs were compared similarity (SIM) with a list of herb names in HerbINFO. A set of herbs which obtained similarity more than or equal to threshold were listed in the smart phone.

5 Experimental Results

5.1 *Herb Names in Thai and English (Including Scientific Names)*

A set of Thai, English, and scientific names of herbs were used as inputs into the system. The result was shown in Table 1. The Corr (Correct) and Ord (Order) mean the correct herb could be found and the order of the correct herbs was shown in the

Table 1 Result from incorrect input of herb names

Herb name	Input name	Incorrect type	Matching name	Corr/Ord	SIM
Holy basil	กะเพา	Misspelling	กะเพรา	Y/1	83.33
Holy basil	กระเพรา	Misspelling	กะเพรา	Y/1	85.71
Sweet annie	โกษฐ์จุฬาลำพา	Different spelling	โกฐจุฬาลำพา	Y/1	84.62
Licorice	กำเฉ้า	Transcription	กำเช่า	Y/1	85.71
Lemon grass	Lemongrass	Different spelling	Lemon grass	Y/1	90.91
Turmeric	Kurcuma longa	Misspelling	Curcuma longa	Y/1	92.31

list, respectively. Red characters present differences between an input name and a matching name.

From the result, some misspellings, different spellings and a transcribed name from Chinese to Thai herb names could be retrieved with the appropriate thresholds if correct herb names were stored in the HerbINFO database. The rank of herbs in the list was usually the first one.

5.2 *Herb Names in Other Languages*

Four languages of herb names, i.e., Chinese (C), Japanese (J), Tamil (T) and Russian (R), were used as inputs into the system. The result was shown in Table 2.

Table 2 Result from multilingual input of herb names translated by Yandex API

English Name	Input language	Output from API	Corr/ Ord	SIM
Ginger	蒜 (C)	Ginger	Y/1	100.00
	薑 (J)	Ginger	Y/1	100.00
	இஞ்சி (T)	Ginger	Y/1	100.00
	Имбирь (R)	Ginger	Y/1	100.00
Aloe Vera	芦荟 (C)	Aloe Vera	Y/1	100.00
	アロエベラ (J)	Aloe Vera	Y/1	100.00
	சோற்றுக்கற்றாழை (T)	Aloe Vera	Y/1	100.00
	Алоэ настоящее (R)	Aloe present	N	50.00
Holy basil	圣罗勒 (C)	Holy basil	Y/1	100.00
	カミメボウキ (J)	Family member bow key	N	-
	துளசி (T)	Basil	Y/2	50.00
	Туласи (R)	Tulasi	Y/3	40.00

From the result, the efficiency of the system depended on the performance of a machine translator and herbs names stored in a database. If the foreign languages names of herbs can be kept into the database, performance of retrieving should be improved, especially herb names in languages, which gain low similarity score and/ or high ranking number of correction, e.g., துளசி (rank = 2) and Туласи (rank = 3).

6 Conclusion and Future Work

In this paper, problems of retrieving information of herbs by their names were addressed. Due to various patterns of writing for herb names, i.e., common names and local names in several languages, synonyms, homographs, misspelling, include names with different spelling. To help visitors retrieve herbal information by any patterns of herb names, a multi-pattern herb name retrieval system was implemented. Using the HerbINFO, a multicultural herbal information system, an approximate string matching technique, and a machine translator, visitors can easily search herbal information by any patterns of herb names. Moreover, contributed herb names from members of the HerbINFO, feedbacks from mobile phone and voting system in the HerbINFO will improve searching in the future. This system will support visitors to find herbal information their need by herb names with their familiar languages and dialects. Multiple machine translators will be used in this system. We left this for our future work.

References

1. Lin K, Friedman C, Finkelstein J (2016) An automated system for retrieving herb-drug interaction related articles from MEDLINE. In: AMIA summits on translational science proceedings 2016, p 140
2. Rivera D, Allkin R, Obón C, Alcaraz F, Verpoorte R, Heinrich M (2014) What is in a name? The need for accurate scientific nomenclature for plants. J Ethnopharmacol 152:393–402
3. Paton A, Allkin R, Belyaeva I, Dauncey E, Govaerts R, Edwards S, Irving J, Leon C, Nic E (2016) Plant name resources: building bridges with users. Botanists 207
4. Lertnattee V, Wangwattana B (2018) Integration of teaching and learning ICT literacy and herbal information in the 21st century. In: The Asian conference on education & international development, pp 39–47
5. Graefe A (2015) Accuracy gains of adding vote expectation surveys to a combined forecast of US presidential election outcomes. Res Politics 2, 2053168015570416
6. Singla N, Garg D (2012) String matching algorithms and their applicability in various applications. Int J Soft Comput Eng 1:218–222

7. Peng T, Li L, Kennedy J (2018) A comparison of techniques for name matching. GSTF J Comput (JoC) 2
8. Campos L, Pedro V, Couto F (2017) Impact of translation on named-entity recognition in radiology texts. Database 2017
9. International tourist arrivals to Thailand 2017. https://www.mots.go.th/more_news.php?cid=465&filename=index last accessed 2018/10/01
10. Yandex API. https://tech.yandex.com/translate/. Last accessed 2018/08/02

Aspect Categorization Using Domain-Trained Word Embedding and Topic Modelling

Omar Mustafa Al-Janabi, Nurul Hashimah Ahamed Hassain Malim and Yu-N Cheah

Abstract Aspect-based sentiment analysis is the most important research topic conducted to extract and categorize aspect-terms from online reviews. Recent efforts have shown that topic modelling is vigorously used for this task. In this paper, we integrated word embedding into collapsed Gibbs sampling in Latent Dirichlet Allocation (LDA). Specifically, the conditional distribution in the topic model is improved using the word embedding model that was trained against (customer review) training dataset. Semantic similarity (cosine measure) was leveraged to distribute the aspect-terms to their related aspect-category cognitively. The experiment was conducted to extract and categorize the aspect terms from SemEval 2014 dataset.

Keywords Aspect-based · Aspect category · Word-to-topic distribution · LDA · Word embedding

1 Introduction

The web has become an essential source of information for the customers to compare and evaluate the products and services. As these vast amounts of content keep enlarging, an automated mechanism is required to transfer human sentiments into a machine-readable content. Many algorithms and approaches have been emerged in the field of natural language processing to make customer's online reviews machine processable. However, opinion mining generally aims to explore opinion-target, opinion expression, target categorization, and of course, the most

O. M. Al-Janabi (✉) · N. H. A. H. Malim · Y.-N. Cheah
School of Computer Science, Universiti Sains Malaysia, Gelugor, Malaysia
e-mail: omar37513@gmail.com

N. H. A. H. Malim
e-mail: nurulhashimah@usm.my

Y.-N. Cheah
e-mail: yncheah@usm.my

Z. Zakaria and R. Ahmad (eds.), *Advances in Electronics Engineering*,
Lecture Notes in Electrical Engineering 619,
https://doi.org/10.1007/978-981-15-1289-6_18

typical task is opinion polarity analysis in online reviews. In a fine-grained analysis, opinion-target extraction and categorize are the most important, yet challenging opinion-mining subtasks. Opinion-target is the aspect/feature of a product in online customer reviews. For instance, in laptop reviews “the battery life seems to be very good, and have had no issues with it”. The aspect or opinion-target is “battery”. Target categorization (it categorizes the same aspect terms into one class) is required to avoid evaluating identical aspect terms differently. By doing so, it would ease the way of assessing the polarity of the extracted aspect-terms.

In this work, we have proposed the continuous word embedding model that has been trained to the customer reviews dataset to avoid using manually labeled knowledge- or dictionary-based. The trained word-embedding later integrated into the collapsed Gibbs sampling in an extended Latent Dirichlet allocation (LDA) that allows side information to be exploited to tune its distributions.

The previous effort has been broken into three main methods to accomplish the task of extracting the aspect-terms from customer reviews. Whereas, topic modeling was used to tackle both aspect extraction and categorization, simultaneously, as next section illustrates.

The paper is structured as follows: briefly stated the main methods that used for aspect extraction in online reviews, the followed section explains our proposed model, and the last section is for result and conclusion.

2 Related Work

The problem of aspect extraction has conducted user’s opinion, particularly focused on the extraction of aspects from online customer reviews [1, 2]. One of the most notable methods that have been used to extract the opinion-target from online reviews is the frequency-based method, which is relied on much more frequently repeated set of nouns and noun phrases over the rest of the vocabulary in sentence review. These nouns have been extracted as aspect-term/opinion-target [3–6].

The second method is the syntax-based/syntactic pattern method. It has been introduced to mitigate the shortage of frequency method of finding aspects, because not all the frequently repeated nouns and noun phrases are aspects. It relies on the syntactical relations between the aspect word and its sentiment. The simple example of syntactical-relation is an adjectival modifier, e.g. ‘fantastic food’, the adjective here is ‘fantastic’, which is modifying the aspect word ‘food’. Further, the syntactic pattern is either hand-crafted rules or automatically generated rules [7–10].

The third method is the supervised machine learning that is being hybridized with either frequency based, or rule-based to extract the aspects, because aspect extraction problem is a special case in information extraction. Supervised machine learning performed well on aspect extraction task when there is comprehensively annotated dataset, and that is not always accessible. Conditional Random Field is one of the most popular supervised machine learning in aspect extraction [11, 12], other supervised methods are [13, 14].

The fourth method is topic modeling, a plethora of probabilistic topic modeling methods conducted the task of aspect-based sentiment analysis ABSA [15]. The critical factor of topic modeling that made it so vigorous, can be used to extract and categorize the similar aspects into their related clusters concurrently [16–20].

A well-known probabilistic method for modeling aspects in documents is Latent Dirichlet Allocation (LDA) [21]. Notable models that are related to our proposed work are; DF-LDA [22], a semi-supervised model that allows the user to use two distribution constraints, must-link, and cannot-link constrain. A must-link constraint is to force two terms to be in the same topic (aspect category). Cannot-link constraints mean that two terms cannot be at the same topic (aspect category). These constraints are expressed with seeds that are manual and should be set up in advance, which lead to the problem of domain-dependent. However, our work is different from DF-LDA, in that, we introduced continuous word embedding to tackle the problem of manually expressed seeds, and we have proposed an extended LDA that accepts external side information.

In [23], SAS model is presented to extract aspect terms by exploiting seeds that are used to discover the specific aspect. While Poria [18] introduced Sentic LDA topic model, which was integrated to common-sense reasoning to extract and categorize the aspects. Their model seems to be domain-dependent, because the word-to-topic distribution in LDA topic models relies on a predefined dataset that includes the semantic-similarity values between each word and the most similar seed-words from the dataset. The new training domain is required to be stored in a separate dataset. However, we have tackled the problem of domain-dependent by integrating the domain-trained word embedded to the collapsed Gibbs sampling in LDA. Also, we adopt the cosine semantic similarity to assure cognitive distribution in word-to-topic distribution in LDA.

3 Model Description

The primary objective of the proposed model is to solve the problem of domain-dependent and improve topic coherency with any domain without an annotated external knowledge or a corpus, such as WordNet. The proposed model is detailed in the following sections.

3.1 Word Embedding

Continuous word embedding model [24] was proposed in this work to generate word embedding from the training dataset (online customer reviews). In other word, it represents a numerical vector for each term in the corpus vocabulary. The learned model has encoded a piece of information about the meaning or concept for each term and the relationship between it and other terms in the corpus. In word embedding, the implementation of each word vector in the corpus has been set to 100 dimensions, the context window is 5, and the epochs 12. Thus, terms that share common context in vector space are positioned in close proximity in word vector. However, the semantic similarity measure (cosine measure) were also trained based on the domain-trained word embedding. The idea of trained semantic similarity measure using word2vector was first proposed in [25].

3.2 Aspect-Term Categorization in Topic Modelling

The topics in topic modeling represented the aspect-category in ABSA that conducted the online customer reviews, where each topic is supposed to contain the most related aspect-terms. In topic modelling like LDA, the extraction of aspect-terms is requiring an additional step to map them to a meaningful aspect-category. Recent effort has empowered the distributions of the word to the topic in LDA by exploiting manually labeled seeds or dictionaries that led to the issue of domain-dependent. This work strives to introduce word embedding to be trained on the customer reviews dataset. The trained word-embedding leverages the distribution of the words (or the aspect in our case) to its related category. An extended LDA [26] was proposed in this work, which explicitly allows side information over the distribution of words.

The proposed scheme of our LDA topic model, was extended by including word embedding knowledge into the conditional distribution in collapsed Gibbs sampling. The generative model is described as follows: for each document $d \in \{1, \ldots, D\}$, a distribution of topics, θ_d, is being sampled from a Dirichlet distribution that is represented by parameter α_d, which is a vector that govern the distribution of topic priors for the document. So far, we are dealing with fine-grained analysis of aspect-based, and each document is referring to a single sentence in a corpus. While, for each sentence $s_d \in \{1, \ldots, s_d\}$, we assume each sentence represents a single aspect, and a distribution of topics, $\emptyset_t$, also sampled from Dirichlet distribution that represented by parameter β, which governs the distribution of aspects to the topic. For each aspect-term (word) $w \in \{w_1, \ldots, w_w\}$ in a sentence s_d, a topic distribution $\emptyset_t$, is biased by topic-seeds, μ_s. These seeds are sampled using hyperparameter ρ, which is a vector that tunes the distribution of word to topic. Nevertheless, the topic-seeds: are the sets of most semantically related terms that being extracted using domain-trained word embedding. Seeds extracted are based on each aspect-category or topic in the trained domain dataset. i.e. topic-seeds for aspect-category ‘ambiance’ in restaurant review dataset is ‘lousy, bistro, café, relax, décor’. Similarly, other aspect categories of topic-seeds are being extracted.

To discover sets of latent topics (aspect categories) used in the corpus, the topic-to-document distribution is being addressed using multinomial distribution, but the word-to-topic distribution is intractable and for this issue, Collapsed Gibbs sampling (CGS) algorithm was introduced. It is used as a conditional distribution to represent word to topic distribution. Our strategy differs from the previous work, in which the topic-seeds are embedded into the conditional distribution to tune parameter β, governs word to topic distribution. CGS gives a complete description of conditional assignment probability of a single-word topic assignment, conditioned on the rest of the model. Given the assignment of latent variables; topic assignment $z_{d,n}$, of each word in the document will be semantically assigned to the related aspect category. Formula (1) shows the conditional distribution of word to topic distribution.

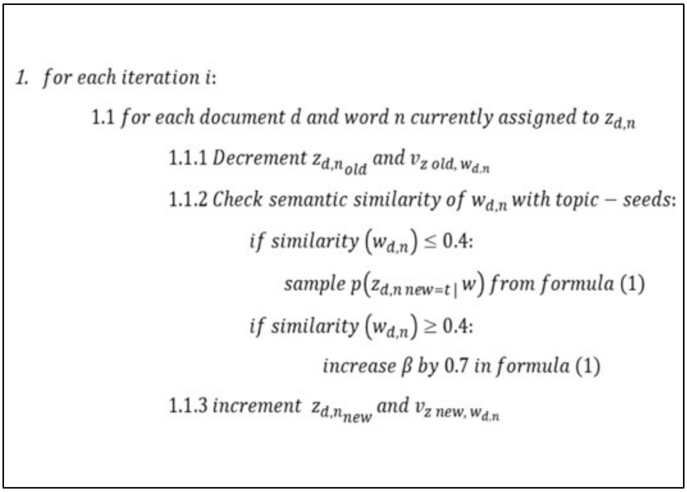
$$p(z_{d,n} = t| - z_{d,n}, w; \alpha, \beta, \rho) \propto \frac{\left(n_t^{(d)} + \alpha\right)}{\left(n_{\cdot}^{(d)} + Ta\right)} * \frac{\left(n_t^{(w)} + \beta\right)}{\left(n_t^{(\cdot)} + w\beta\right)} * \frac{\left(n_s^{(d)} + \rho\right)}{\left(n_t^{(\cdot)} + s\rho\right)} \quad (1)$$

where, $n_t^{(d)}$ is the number of times an aspect-term from document d, has been assigned to aspect-category, $n_t^{(w)}$ is the number of times aspect-term has been assigned to aspect-category, and $n_s^{(d)}$ is the number of times a seed from topic-seeds (that have been extracted using word embedding) in document d, has been assigned to aspect-category.

Practically, after the incorporation of the topic-seeds into the conditional distribution, and the multinomial distribution of the document to the topic. LDA collapsed Gibbs sampling algorithm is then modified by adding semantic similarity measure to tune the word-to-topic distribution, which is related to parameter β. The intuition of proposing semantic similarity measure is to assess the similarity between $w_{d,n}$ and the extracted topic-seeds.

3.3 An Improved CGS

The novelty of this work uses a measure of similarity among words based on the domain-trained word embedding, that have been stated earlier at Sect. (3.1). Pseudocode (1) shows the improved CGS, which has leveraged the semantic similarity measure to draw the word to topic cognitively. For each $d \in \{1, \ldots, D\}$ document, for every $w_{d,n}$ in the currently assigned $z_{d,n}$ aspect-category, within i number of iterations, decrement the variable associated with the topic assignment. The domain-trained word embedding has cognitively improved the drawing of the aspect-terms; the extracted topic-seeds guided the distribution of aspect-terms, and the modified CGS has leveraged the semantic similarity measure that cognitively assures the coherency of the topics based on the customized threshold of similarity between current word and topic-seeds. The lowest threshold for similarity is 0.4, unlikely the current $w_{d,n}$ will be sampled to the $z_{d,n}$, thus, β is set to 0.1. Otherwise, if the semantic similarity of the current $w_{d,n}$ is ≥ 0.4, it is most likely to be sampled to the $z_{d,n}$, if so, increase the value of β to 0.7, to assign the current word to its aspect-category, and increment the topic assignment again. The long-term iterations in the modified CGS has produced more coherent topics.



```
1. for each iteration i:
        1.1 for each document d and word n currently assigned to $z_{d,n}$
            1.1.1 Decrement $z_{d,n_{old}}$ and $v_{z\ old,\ w_{d,n}}$
            1.1.2 Check semantic similarity of $w_{d,n}$ with topic − seeds:
                if similarity $(w_{d,n}) \leq 0.4$:
                    sample $p(z_{d,n\ new=t\,|}\ w)$ from formula (1)
                if similarity $(w_{d,n}) \geq 0.4$:
                        increase $\beta$ by 0.7 in formula (1)
            1.1.3 increment $z_{d,n_{new}}$ and $v_{z\ new,\ w_{d,n}}$
```

Pseudocode (1): Collapsed Gibbs sampling

4 Results and Conclusion

This section discusses the performance and what has been concluded from the proposed model. However, a benchmark dataset, SemEval-2014 was used to assess the performance of our model. It also being used to train the word embedding model in this work. In this dataset, there are two different costumer-reviews: Restaurant, and laptop domain. As for the proposed model is an unsupervised probabilistic model, the evaluation is mainly relied on the Restaurant domain dataset (because the labels of this dataset are available). Thus, the aspect-terms are being distributed over five aspect-categories: food, service, price, ambiance, and anecdotes.

The values for the proposed model are: 0.1 for α, that is for the whole experiment, the value for β varies based on the semantic similarity as mentioned in Sect. (3.3), while ρ is set to 15. There are 5 topics/aspect-categories, and 3000 iterations.

Two models were chosen to be compared with our model; [27], and Sentic LDA, as shown in the Table 1.

Table 1 Topic model's performance

Models	Precision (%)	Recall (%)
Kiritchenko [27]	91.04	86.24
Sentic LDA	92.12	88.25
Our model	84.03	81.39

To conclude the proposed work, the domain-trained word embedding solve the problem of domain-dependent, where the word embedding could be used to perform any domain analysis. The incorporation of semantic similarity (cosine measure) with collapsed Gibbs sampling is an alternative solution for manually labeled seed or an external knowledge like WordNet, which maintains the topic coherency. However, the performance of our model lower than the previously proposed models because in their models an annotated seed-sets were used to distribute the aspect-terms and that time-inefficient. One limitation of our work, is that is it slower in convergence than the standard model due to the iteration of the model has to go through the model two times. the issue of slow convergence will be tackled in future work using an online LDA.

Acknowledgements This work is supported by University Science Malaysia research grant (1001/PKOMP/8014002—Detection of User Opinion and Emotion in Online Product Reviews).

References

1. Chen Z, Mukherjee A, Liu B (2014) Aspect extraction with automated prior knowledge learning. In: 52nd annual meeting of the association for computational linguistics, ACL
2. Liu K, Xu L, Zhao J (2012) Opinion target extraction using word-based translation model. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning
3. Blair-Goldensohn S, Neylon T, Hannan K, Reis GA, McDonald R, Reynar J (2008) Building a sentiment summarizer for local service reviews. Work NLP Inf Explos Era
4. Meng X, Wang H (2009) Mining user reviews: from specification to summarization. In: Proceedings of the ACL-IJCNLP 2009 conference short papers. ACM
5. Wu Y, Zhang Q, Huang X, Wu L (2009) Phrase dependency parsing for opinion mining. In: Proceedings of the 2009 conference on empirical methods in natural language processing: Volume 3—EMNLP
6. Zhang W, Xu H, Wan W (2012) Weakness finder: find product weakness from Chinese reviews by using aspects based sentiment analysis. In: Expert systems with applications. Elsevier
7. Asghar MZ, Khan A, Zahra SR, Ahmad S, Kundi FM (2017) Aspect-based opinion mining framework using heuristic patterns. Clust Comput (Springer)
8. Liu Q, Gao Z, Liu B, Zhang Y (2015) Automated rule selection for aspect extraction in opinion mining. In: IJCAI international joint conference on artificial intelligence
9. Poria S, Cambria E, Ku LW, Gui C, Gelbukh A (2014) A rule-based approach to aspect extraction from product reviews. In: Proceedings of the second workshop on natural language processing for social media
10. Rana TA, Cheah YNN (2017) A two-fold rule-based model for aspect extraction. In: Expert systems with applications, vol 89. Elsevier

11. Luo H, Li T, Liu B, Wang B, Unger H (2018) Improving aspect term extraction with bidirectional dependency tree representation. Arxiv.org
12. Rubtsova Y, Koshelnikov S (2015) Aspect extraction from reviews using conditional random fields. Springer link
13. Poria S, Cambria E, Gelbukh A (2016) Aspect extraction for opinion mining with a deep convolutional neural network. In: Knowledge-based system. Elsevier
14. Yu J, Zha Z, Wang M, Wang K, Chua T (2011) Domain-assisted product aspect hierarchy generation : towards hierarchical organization of unstructured consumer reviews. Comput. Linguist. MIT Press
15. Rana TA, Cheah YN, Letchmunan S (2016) Topic modeling in sentiment analysis: a systematic review. J ICT Res Appl
16. Bagheri A, Saraee M, De Jong F (2014) ADM-LDA: an aspect detection model based on topic modelling using the structure of review sentences. Artic J Inf Sci (Sage)
17. Jiang W, Pan H (2016) Aspect extraction in product reviews via an improved unsupervised method. In: 2015 2nd international symposium on dependable computing and internet of things (DCIT)
18. Poria S, Chaturvedi I, Cambria E, Bisio F (2016) Sentic LDA: improving on LDA with semantic similarity for aspect-based sentiment analysis. In: 2016 international joint conference on neural networks (IJCNN)
19. Shams M, Baraani-Dastjerdi A (2017) Enriched LDA (ELDA): combination of latent Dirichlet allocation with word co-occurrence analysis for aspect extraction. In: Expert systems with applications (Elsevier)
20. Zhang Y, Tang F, Barolli L, Yang Y, Xu W (2017) Jointly modeling multi-grain aspects and opinions for large-scale online review. In: Proceedings—International conference on advanced information networking and applications IEEE
21. Blei DM, Ng AY, Jordan MI, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res
22. Andrzejewski D, Zhu X, Craven M (2009) Incorporating domain knowledge into topic modeling via Dirichlet Forest priors. In: Proceedings of the 26th annual International Conference on Machine Learning—ICML
23. Mukherjee A, Liu B (2012) Aspect extraction through semi-supervised modeling. Association for Computational Linguistics
24. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. Arxiv
25. Charlet D, Damnati G (2017) SimBow at SemEval-2017 task 3: soft-cosine semantic similarity between questions for community question answering. In: Conference: Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)
26. Petterson J, Smola A, Caetano T (2011) Word features for latent Dirichlet allocation. Adv Neural Inf Process Syst
27. Kiritchenko S, Zhu X, Cherry C, Mohammad S (2014) NRC-Canada-2014: detecting aspects and sentiment in customer reviews. In: Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)

Experimental Investigations of the Effect of Temperature on Power in a Combined Photovoltaic Cell and Thermo-Electric

Ahmed O. Mohamedzain, Sunil Govinda, H. S. Chua and Ammar A. M. Al-Talib

Abstract In this paper, a combination of thermoelectric (TEG) and low power mono-crystalline photovoltaic cell (PV) has been designed and fabricated. The experiment was to investigate and analyze the effect of temperature on PV cell and to investigate the output power generated by TEG under two (2) different levels of irradiance. The temperature and output power of their characteristics within the system set up are determined. Experimental investigation tests have been carried out under solar simulator using halogen light as a light source. Analysis on combined system Photovoltaic Cell/Thermoelectric Generator (PV/TEG) has been conducted based on the different temperature effects at irradiance of 1000 and 750 W/m^2. At irradiance of 750 W/m^2, the maximum temperature at the PV cell hot side was 61 °C, while the output power of TEG and PV cell were 0.0039 W and 2.49 W, respectively. Results have shown that the efficiency of PV cell was improved by 1.4% when PV/TEG system was applied; maximum efficiency of PV cell at irradiance of 750 W/m^2 is 12.18% and when the temperature was 30 °C.

Keywords Photovoltaic · Thermoelectric generator · Photovoltaic cell/thermoelectric generator

A. O. Mohamedzain · S. Govinda (✉)
Department of Electrical and Electronic Engineering, UCSI University,
Kuala Lumpur, Malaysia
e-mail: sunil@ucsiuniversity.edu.my

A. O. Mohamedzain
e-mail: ahmedalzain90@hotmail.com

H. S. Chua
School of Engineering, KDU University College, Shah Alam, Malaysia
e-mail: hs.chua@kdu.edu.my

A. A. M. Al-Talib
Department of Mechanical Engineering, UCSI University, Kuala Lumpur, Malaysia

Z. Zakaria and R. Ahmad (eds.), *Advances in Electronics Engineering*,
Lecture Notes in Electrical Engineering 619,
https://doi.org/10.1007/978-981-15-1289-6_19

1 Introduction

Renewable energy opposes a substantial part nowadays in research and demanding in the market. Solar photovoltaic cell (PV) technology is renewable and considered as one of the most lessening GHG emissions compared with other sources of energy [1]. PV system is greenhouse gas emissions, scalable, reliable, clean, silent, long term, and environmental friendly electricity system for a sustainable future. Recently, researchers have more interest in different types of solar hybrid system, such as photovoltaic Cell/Thermoelectric Generator (PV/TEG), Photovoltaic Cell/Solar Collector (PV/STC), and Photovoltaic Cell/Thermoelectric Generator/Solar Collector (PV/TEG/STC) [2]. TEG operates based on Seebeck effect which is the most common aspect for thermoelectric theory [3]. It is harvesting heat and converts thermal energy into electricity.

Irradiance and temperature represent the key factors that affect PV cell [4]. The most suitable temperature on PV cell to provide a maximum efficiency power output is between 15 and 35 °C [5]. Siecker, Kusakana, & Numbi, 2017 presented that under Standard Test Condition, the operational efficiency of the PV cell is decreased by about 0.40–0.50% for each degree rise in temperature. Therefore, TEG might be applied at the PV cell to harvest the heat and turn it into electricity. Simultaneously, temperature on the PV cell would have noticeable drop resulting an improvement in the sufficiency of the power output.

PV cell efficiency lies between 4% and 17% [6] depending critically on the PV cell type. Therefore significant amount of incident solar energy is instantly converted to heat, which results in increasing of the operated temperature of the PV cell. PV cell cooling methods could be addressed as passive method and active method. Passive cooling methods do not require external power, while active cooling methods are operated by motor pumps or electrical fans, which require external power [7]. Experimental studies on progressively reducing the desired temperature of the PV cell were carried out by many researches on the alternative methods, such as passive cooling system with cotton wick structure [6], hybrid photovoltaic-Thermal PV/T [8], film cooling module [9], hybrid solar PV/Thermal system cooled by water [10], and PV cell cooled by transparent coating (photonic crystal cooling) [10]. There are only a few studies on the effect of TEG on the PV cell with not much of experiment results. Review on recent advance of photovoltaic system has been done by [11] concluded that photovoltaic-thermal (PV/T) is one of the most technical approaches for future energy challenges. Modelling results [12] demonstrated that the temperature of PV cell is reduced when TEG was attached to the backside of PV cell, and the excess heat was able to convert extra energy [13], which results in improving the performance of the system, thus, experimental investigations in these temperature reductions are to be carried out and verified.

The objective of these experiments were to determine the effect of TEG reflected to the reaction temperature and power output on the PV cell and TEG. The full hardware setting up was explained in this paper. As a preliminary studies, the influence of temperature conditions on TEG and PV cell performance was investigated, and the minimum operating conditions were determined.

1.1 Photovoltaic Model (PV Cell)

The electrical characteristic of PV cell is shown in Table 1. the PV cell consists of current source IL (light current), the irradiance and temperature are represented by the series resistance and shunt resistance (Rsh, Rs), and diode as shown in Fig. 2. The I-V curve of PV cell shown in Fig. 1 was drawn using Matlab simulink according to the characteristics of the PV module in Table 1; the characteristic of the PV cell is nonlinear so that the power is changing according to the variation of the irradiance and Temperature. Equations (1) and (2) determine the diode current and the thermal voltage of one cell diode for PV cell model where I_o represents the reverse saturation current; I_d represents diode current; V_d represents voltage across diode; V_T represents a thermal voltage; q is elementary charge; K is Boltzmann's constant; T is the absolute temperature.

$$I_d = I_o\left[\exp\left(\frac{V_d}{V_T}\right) - 1\right] \tag{1}$$

$$I_d V_T = \frac{KT}{q} \tag{2}$$

Table 1 Electrical characteristics of PV cell

Item	Specification
Max power (Pmax)	4.5 W
Number of cells	12
Max voltage (Vmax)	6 V
Max current (Imax)	750 MA
Open Circuit Voltage (Voc)	7.2 V
Short Circuit Current (Isc)	850 MA
Size	165 * 165 * 3 MM

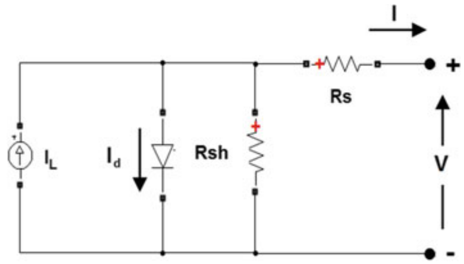


Fig. 1 Electrical circuit of PV cell with one cell diode

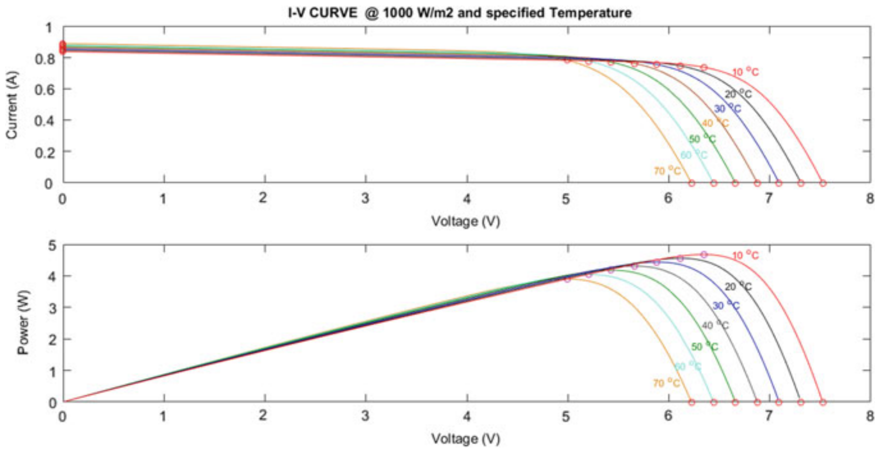


Fig. 2 I-V and P-V curve of PV cell under different temperatures

1.2 Thermoelectric Model (TEG)

Electricity can be produced directly from the differences of temperature at specified two junctions of semiconductor metal. Equation 3 [14] represents the Seebeck concept, where, T1 and T2 are the temperatures at two junctions. SA and SB represent the Seebeck coefficients.

$$V = \int_{T1}^{T2} (S_B(T) - S_B(T))dt \tag{3}$$

Scientifically, based on the datasheet of standard TEG module (TELBP1-12656-0.45), the TEG module was typically constructed from Lead Tin Tellurium and Bismuth Tellurium with the maximum operation temperature up to 200–360 °C. At any values during the temperature difference between hot and cold sides, TEG would generate potential DC power. The higher the temperature difference, the higher output power. Therefore, the output DC power could reach up to 247 W at 350 °C at hot side, while its 30 °C at cold side as shown in Table 2.

2 Methods

2.1 Hybrid Model (PV/TEG) Fabrication

As shown in Fig. 3, two halogen lights were used as solar simulator. At the distance of 37 cm from the luminous source, the irradiance was 750 W/m^2. At the distance of 22.5 cm from the luminous source the irradiance was 1000 W/m^2. The PV cell

Table 2 Specification of thermoelectric generator

Item	Specification
Hot Side Temperature/Th (°C)	350
Cold Side Temperature/Tc (°C)	30
Open Circuit Voltage (V)	9.2
Matched Load Resistance (ohms)	0.97
Matched load output voltage (V)	4.6
Matched load output current (A)	4.7
Matched load output power (W)	21.7
Heat flow across the module(W)	247
Heat flow density (W cm^{-2})	7.9
AC Resistance (ohms) Measured under 27 °C at 1000 Hz	0.42–0.52
Size	56 * 56 * mm

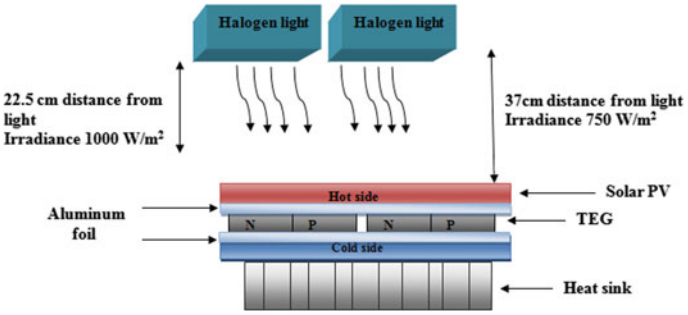


Fig. 3 PV/TEG module

was properly placed on top of the model and the TEGs were attached at the bottom of PV cell. TEGs would instantly harvest the heat from the PV cell and convert it to power. Subsequently, the heat-sink that was properly attached at the bottom of the module would act efficiently as a cooler to absorb heat from the hot side of the TEGs. Four pieces of TEG modules (TELBP1-12656-0.45) have been attached to the model, which was connected in series-parallel electrically (2//2). The aluminum foil was attached to the back of PV cell to improve heat transfer from the PV cell to TEGs.

2.2 *Hardware Preparation*

Four (4) TEGs (TELBP1-12656-0.45) were properly connected in parallel series as shown in Fig. 4 to maximize the generated output current and they were attached to the backside of the mono-crystalline PV cell (4.5 W). The heat at the PV cell from

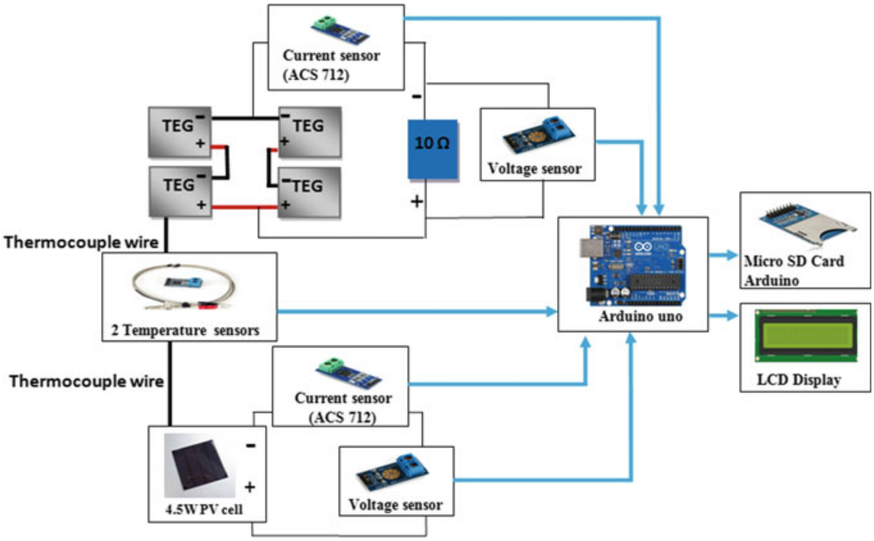


Fig. 4 Schematic of TEGs connection and PV cell

the back side was transferred to the hot side of TEG, while the cold side of TEG is attached to a heat sink. This phenomenon set up is called Seebeck effect and forms an electricity effect. Arduino Uno microcontroller has been used in the system to measure and record the temperature and output power of the PV cell and TEG into memory card. A Thermocouple sensor K-type and MAX6675 driver are used to measure the temperature at each side of the model. Liquid Crystal Display (LCD) (I2C 2×16) was connected to the microcontroller to display temperature values in Celsius unit, which was instantly updated and accurately recorded every 5 s. Two (2) voltage sensors and two (2) current sensors (ACS712) were used to accurately measure the output voltage and output current of PV cell and TEGs. Irradiance measurements were carried out at deferent heights from the light source, using solar meter (TES 1333R).

3 Experiment Setup

Two (2) PV cells were used to be tested and compared in terms of output power efficiency. Arduino UNO was used as data acquisition to measure the voltage, current, and temperature. Figure 5 displays the custom fabricated hardware, which was designed to support the TEGs and the whole completed setup of the experiments. The system was studied for 2 h until a constant power was observed. The Two (2) TEGs were connected in series. The two (2) sets of TEGs were connected

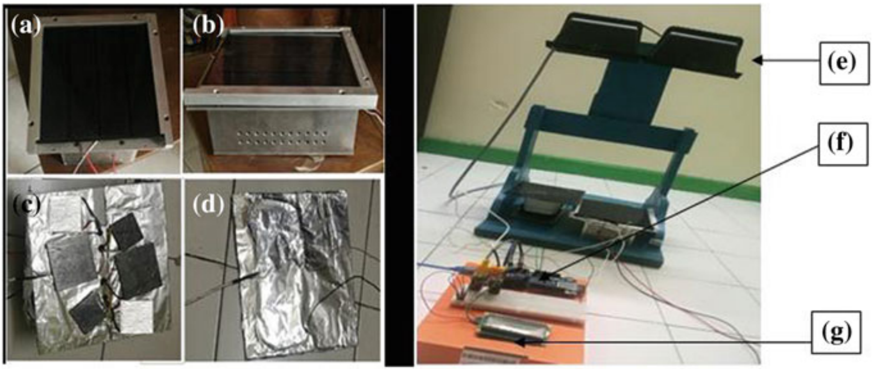


Fig. 5 **a** Top view of hybrid system, **b** Side view of hybrid system, **c** Connection of TEGs, **d** PV cell attached with aluminum foil, **e** Solar simulator halogen light, **f** Voltage, current, and temperature sensors connected to Arduino Uno, **g** LCD Display

in parallel to increase the current collection. The TEGs were attached to the bottom of PV cell to increase the efficiency of the power output by harvesting the temperature of PV cell and converting it to power.

3.1 Non-uniformity Calculation

Non-uniformity (SNU) over the specified test area is the most important parameter in solar simulator. According to the American Society for Testing and Materials (ASTM) and the International Electro technical Commission (IEC) standards, if the SUN is less than 5% the area of a solar simulator is standardized as effective test area. SNU was evaluated [2] at two different heights. SUN is defined by Eq. 3, where E_{max} is the maximum measured intensity over the test area, and E_{min} is the minimum measured intensity over test area.

$$SNU = \frac{E_{\max} - E_{\min}}{E_{\max} + E_{\max}} \times 100\% \tag{4}$$

At heights of 22.5 cm and 37 cm, solar irradiance was 1000 W/m^2 and 750 W/m^2, SNU was 3.35% and 3.8%, respectively, which was standardized as effective test area.

4 Results and Discussion

Hardware results were divided into two parts. The first part is the results of TEGs, which illustrates the effect of temperature on TEGs. The second part is the results of combined system (PV/TEG), which presents the reading of output power of PV cell under different levels of temperature.

4.1 TEGs Results

Table 3 shows the output power of TEGs when the irradiance was 1000 W/m^2. Output power of TEG increased significantly when the temperature increased at the hot and cold sides of the TEGs. 82 °C has been observed at the hot side of TEGs when the irradiance was 1000 W/m^2. At the same time, the power generated from the TEGs was 0.0192 W. as shown in Table 4, temperatures dropped to 61 °C when the irradiance was 750 W/m^2, which results in decreasing the output power of the TEGs to 0.0039 W. Voltage drops explained how the overall power was decreased when temperature of TEGs decreased. The output voltage of the TEGs was 0.17 V when temperature in between 50 and 52 °C, while output voltage of

Table 3 Output power of TEGs at 1000 W/m^2

Temperature at hot side (Th) (°C)	Temperature at cold side (Tc) (°C)	Output power of TEGs (W)
30	26	0.000169
44	30	0.00242
52	33	0.00496
60	38	0.0078
68	43	0.0138
76	50	0.01739
82	58.5	0.0192

Table 4 Output power of TEGs at 750 W/m^2

Temperature at hot side (Th) (°C)	Temperature at cold side (Tc) (°C)	Output power of TEGs (W)
30	26	0.00018
36	28.5	0.00054
40	30	0.0012
44	31.5	0.00182
52	34	0.00256
56	36.5	0.003298
61	40	0.003933

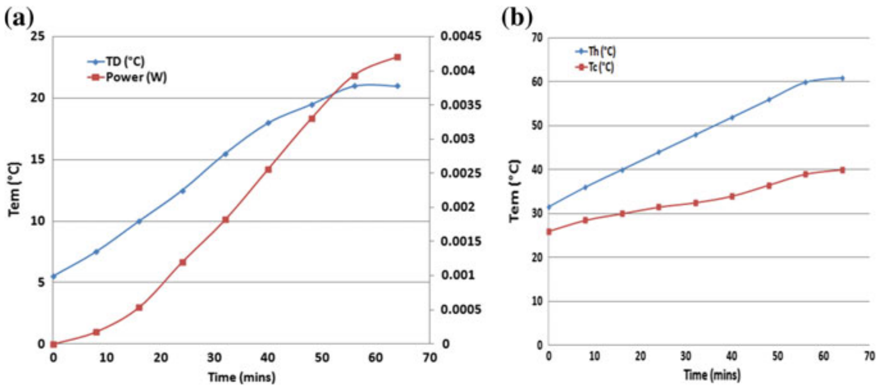


Fig. 6 **a** Output Power of TEGs Compared with Difference in Temperature by Attaching Aluminum Foil at 750 W/m^2; **b** Temperature at Cold and Hot Side by Attaching Aluminum Foil, at 750 W/m^2

TEGs was 0.14 V when temperature was in between 55 and 52 °C. It was observed that the power generated of the TEGs was 0.0192 W when the temperature was 82 °C, while the irradiance was at approximately 1000 W/m^2. The maximum temperature of TEGs from the hot side was 65 °C when the irradiance was at approximately 750 W/m^2. It took 13 min to reach a nearly constant 0.002 W when the irradiance was 1000 W/m^2, while it took 30 min at irradiance of 750 W/m^2. Based on the experiment, outside factors such as wind have strong effects on the performance of the hybrid model. From Fig. 6b, we observed that the temperature took 70 min to reach 65 °C at the hot side of the TEGs when the irradiance was 750 W/m^2, and it took only 25 min to reach 65 °C when irradiance was 1000 W/m^2 as shown in Fig. 7b.

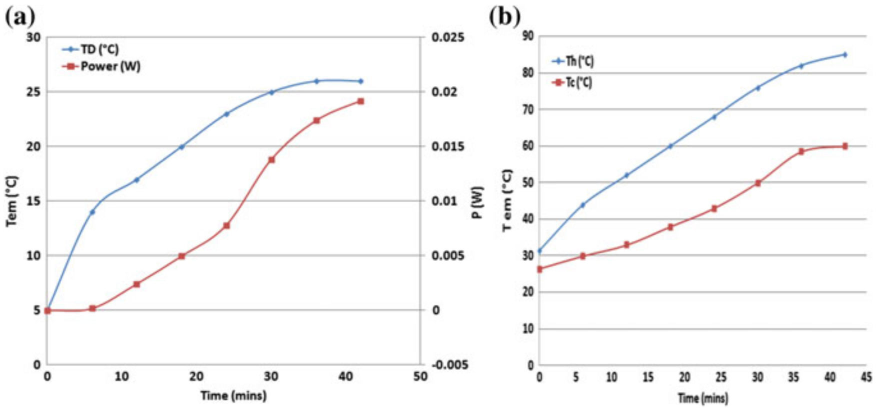


Fig. 7 **a** Output Power of TEGs by Attaching Aluminum Foil at 1000 W/m^2; **b** Temperatures at Cold and Hot Side by Attaching Aluminum Foil at1000 W/m^2

Temperature difference (TD) between PV cells using TEGs and without using TEGs, reached 4.5 °C when the irradiance was 750 W/m^2.

4.2 Combined System Results

Two PV cells were evaluated in this experiment. One of the PV cells was designed with applying PV/TEG system, the second PV cell was designed without applying PV/TEG. Figure 8 presents the efficiency of PV cell with and without applying TEGs. The efficiency was calculated from the output power measurements of the PV cell. Therefore, the measured output power of the PV cell was divided by the area of the PV cell (165 * 165 * 3 mm) (W/m^2). Then, the values of the power in W/m^2 were divided by the measured irradiance which was 750 W/m^2. Then, it was multiplied by 100 to determine the efficiency of PV cell at different temperature conditions. It took minutes for the output power of the PV cell to reach 2.01 W when the irradiance was 750 W/m^2, while the difference power PD between PV cells with and without applying TEGs was 0.255. Simultaneously, the difference temperature TD between PV cells with and without applying TEGs was 3 °C. The improved power of the PV cell was 0.362 W by applying TEGs and aluminum foil. The time taken was two hours for this experiment. Average power difference between two PV panels of PD was 0.26 W at irradiance of 750 W/m^{2,} which improved the efficiency by average of 1.4%. Maximum efficiency of PV cell has been observed at operating temperature of 30 °C.

5 Conclusion

The combined system PV/TEG has been tested under solar simulator using halogen light as a light source irradiance and temperature data of the combined system PV/TEG has been analyzed. Results have proven that decreasing the temperature leads to higher output power from PV cell. Maximum improved power of PV cell was 0.362 W. At irradiance of 750 W/m^2, after 30 min, temperature of PV cell in combined system was 45 °C, while temperature of PV cell without applying combined system was 53 °C at the same time. At irradiance of 1000 W/m^2, PV cell temperature reached up to 82 °C, in addition, efficiency of PV cell improved by 1.4%. Maximum power generated by the TEG model was 0.0192 W; which lead to a conclusion that TEG contribute to the output power of PV panel was insignificant. However, TEGs attached to heat-sink contributes in decreasing the temperature of the PV cell, which leads to increase the lifetime of the PV cell. There are many countries already looking toward a waste to energy (WtE) using incineration [13] and pyrolysis [14]. These processes are associated with high temperatures ranging

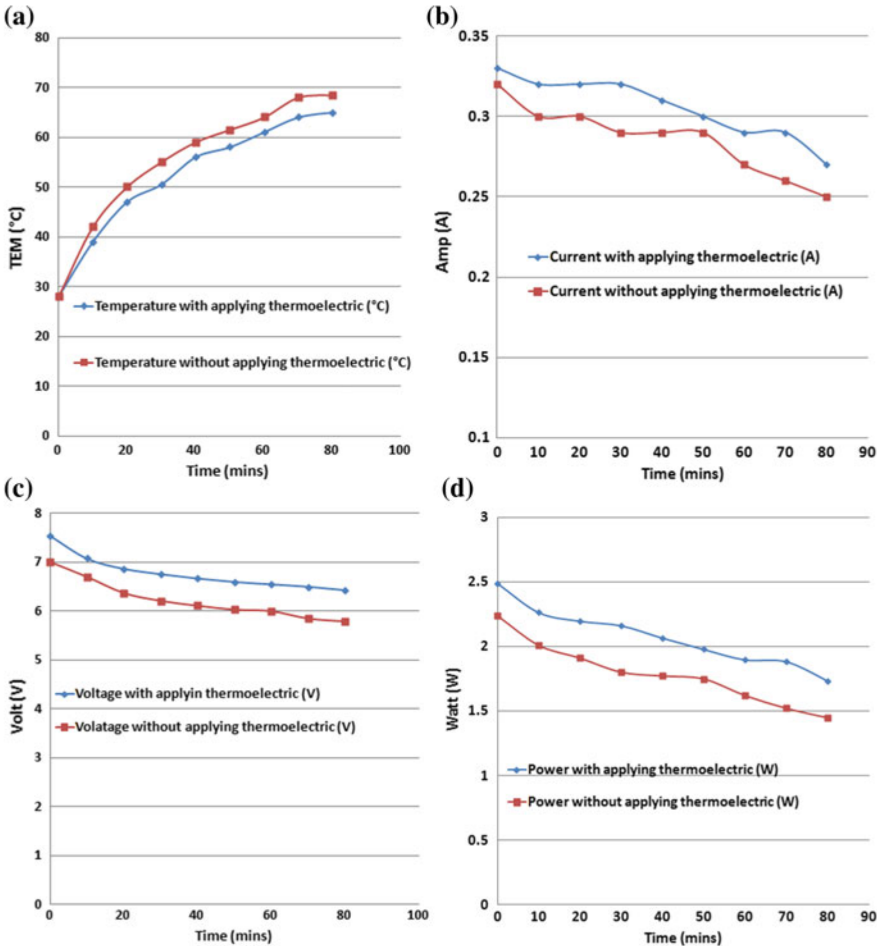


Fig. 8 **a** Output voltage of PV cell with and without applying PV/TEG at 750 W/m^2, **b** Output power of PV cell, with and without applying PV/TEG at 750 W/m^2, **c** Temperature of PV cell with and without applying PV/TEG at 750 W/m^2, **d** Output current of PV cell with and without applying PV/TEG at 750 W/m^2

from 350 to 600 °C. A highest Calcium/Manganese (CMO) TEG module works up to 800 °C with little degradation over 50 years [2]. The size is 65 mm × 65 mm square millimeter and able to produce 9 W when hot side surface at 432 °C, cold side surface at 40 °C. Thus, a higher temperature is required for a higher power output.

Acknowledgements The authors acknowledge the financial support by the UCSI University. And we would like to thank Ir. Associate Professor Dr. Rodney Tan Hean Gay, for his kind support. We would like to acknowledge the equipment support by KDU University College, Glenmarie.

References

1. Abbasov and Elnur (2016) Sustainable solution for increasing the share of solar photovoltaic usages on residential houses in Azerbaijan. Environ Res Eng Manag 71(714):11–18
2. Cotfas DT, Cotfas PA, Floroian L, Floroian DI (2017) Study of combined photovoltaic cell/thermoelectric element/solar collector in medium concentrated light. In: IEEE, pp 747–752
3. Chandrasekar M, Suresh S, Senthilkumar T, Ganesh Karthikeyan M (2013) Passive cooling of standalone flat PV module with cotton wick structures. Energy Convers Manag 71:43–50
4. Hasanuzzaman M, Malek ABMA, Islam MM, Pandey AK, Rahim NA (2016) Global advancement of cooling technologies for PV systems: a review. Sol Energy 137:25–45
5. Joshi AS, Tiwari A (2007) Energy and exergy efficiencies of a hybrid photovoltaic-thermal (PV/T) air collector. Renew Energy 32(13):2223–2241
6. Abdellatif O (2013) Experimental investigation of different cooling methods for photovoltaic module. In: 11th Int Energy Convers Eng Conf, pp 1–7
7. Siecker J, Kusakana K, Numbi BP (2017) A review of solar photovoltaic systems cooling technologies. Renew Sustain Energy Rev 79(May):192–203
8. Pandey AK, Tyagi VV, Selvaraj JA, Rahim NA, Tyagi SK (2016) Recent advances in solar photovoltaic systems for emerging trends and advanced applications. Renew Sustain Energy Rev 53:859–884
9. Cotfas PA, Cotfas DT, Machidon OM (2016) Modelling and PSPICE simulation of a photovoltaic/thermoelectric system. In: 2016 IEEE 22nd Int Symp Des Technol Electron Packag SIITME 2016, pp 179–183
10. Huseynov A, Abbasov E, Salamov O, Salmanova F, and K. technologijos Universitetas (2015) Hybrid solar-wind installation prospects for hot water and heating supply of private homes on the apsheron peninsula of the Republic of Azerbaijan. TT - Karšto vandens ruošimo ir šildymo nuosavuose namuose, naudojant hibridinį saulės ir vėjo energijos įre. Environ Res Eng Manag 71(3):36–48
11. Daud MMM, Nor NBM, Ibrahim T (2012) Novel hybrid photovoltaic and thermoelectric panel. In: 2012 IEEE International Power Engineering and Optimization Conference PEOCO 2012—Conf Proc, no. June, pp 269–274
12. Irwan YM, Leow WZ, Irwanto M, Fareq M, Amelia AR, Gomesh N, Safwati I (2015) Indoor test performance of PV panel through water cooling method, vol 79, Nov 2015, pp 604–611 (Elsevier B.V.)
13. Chua HS, Bashir MJK, Tan KT, Chua HS (2018) A sustainable pyrolysis technology for the treatment of municipal solid waste in Malaysia. Paper presented at the conference proceedings of 2018 International Conference on Environment (ICENV 2018) for Journal of Environmental Chemical Engineering, Dec 2018
14. Huang Shen C, Bashir MJK, Tan KT, Joceyln LPG, Albert FYC (2018) Design and implementation of a laboratory-scale pyrolysis combustor for biomass conversion. Sci Int (Lahore) 30(1):81–84

Impact of Different Annealing Processes on the Performance of Nylon-Based Artificial Muscles for the Use in Robotic Prosthetic Limbs

Nurul Anis Atikah, Yeng Weng Leong and Adzly Anuar

Abstract Artificial muscles made of nylon fishing strings can be actuated using heating and cooling or induced electrically or thermally. These nylon fishing strings show promise in replacing the bulky and expensive actuators that are typically used in robotic prosthetic, as they are lightweight and relatively low in cost. The lifespan and performance of these artificial muscles need to be identified for further development in the use of prostheses and artificial muscles. In this research, we conducted experiments, where the nylon strings that have been coils into Super-coiled polymer (SCP) were tested after going through four different conditions of annealing process. these strings are then tested in lab-based rig to be checked for durability and performance. From the result, we discover that using Slow heating slow cooling (SHSC) annealing process makes the artificial muscles become more durable and while the string undergoes Fast heating fast cooling (FHFC), it can exert higher force comparing to other approaches. This result proves to be useful in deciding the most suitable process to prepare the string to obtain the best performance for the use in robotic prosthetic.

Keywords Artificial muscles · Nylon fishing strings · Annealing process

1 Introduction

Technologies revolutions in microelectronics, artificial intelligence, and material science have resulted in exceedingly discovery in robotics, exoskeletons. And prosthetics [1]. Semiautonomous humanoid robot such as ASIMO (Advance Step in

N. A. Atikah (✉) · Y. W. Leong
Department of Electronics and Communication Engineering,
UNITEN, Jalan Ikram-Uniten, 43000 Kajang, Selangor, Malaysia
e-mail: anisatikah17@gmail.com

A. Anuar
Department of Mechanical Engineering, UNITEN, Jalan Ikram-Uniten,
43000 Kajang, Selangor, Malaysia

Z. Zakaria and R. Ahmad (eds.), *Advances in Electronics Engineering*,
Lecture Notes in Electrical Engineering 619,
https://doi.org/10.1007/978-981-15-1289-6_20

Innovative Mobility) that was designed by Honda were envisioning to take care of the aging population of humans. Exoskeletons also have been commercially available and have been designed to help to improve and aid disabled people to walk and carry out their everyday task [2]. Prosthetics that mimic human-like artificial muscles have been discovered by various researchers [3]. The most well-known artificial muscles are the McKibben Artificial Muscles [4]. They provide fast response and have high power densities. However, this actuator is driven by different types of electric motors that make these actuators bulky, heavy, stiff, and noisy, and were less accepted by the user [5]. Thus, actuators, made of various kinds of materials, such as dielectric elastomers [6], carbon nanotubes [7], and shape memory polymers (SMP) were emerged [8].

The main issues among these actuators are in terms of their performance, scalability, and cost of the materials used to make the actuators. A researcher by the name Haines [9] successfully discovered that twisted nylon fishing lines or Super-coiled polymer (SCP) can match or exceed the performance of mammalian skeletal muscles to deliver millions of reversible contractions and over 20% tensile stroke, while rapidly lifting heavy loads [10]. This SCP can produce larger strain and faster relaxation speed compared to other materials such as Shape memory alloy (SMA). These SCP actuators can be activated thermally or electrically, where high temperature from hot air gun or hair dryer can make the actuator move. Applying Joule heating to activate the SCP, that has been coated with conductive silver paint, allows electricity flow and makes the SCP actuated [11]. With this discovery, simple and cheap solutions can be used in producing artificial muscles for prosthetic purposes. Typically, to fabricate the artificial muscle, nylon strings will be twisted or braided into coils as it would be more compact and easier to be installed in robotic prosthetic limb. Also, in general, it would enhance the strength of the SCP [12]. From previous research, different kinds, and size of nylon fishing lines gives different times to reach their respective maximum tensile force [13]. Therefore, preparation of fabrication of SCP can impact the performance of artificial muscles. The twisted coil, were suggest to undergoes annealing process, where it can increase the life span of the SCP [14]. In this research, we are investigating the effect of using different approach on annealing process on the performance of the SCP, where the rate of heating and cooling are changed. These SCPs were then tested and the result, in terms of durability and maximum tensile force exerted, were tabulated.

2 Methodology

2.1 Design of Testbed

Designing the testbed is very important and needs to be designed by considering several aspects, such as the material, size of the testbed, and its relative position during the experiment. Material used in making this testbed is very important because it needs to withstand the force of the Super coiled polymer during the experiment. From previous research a box made of Perspex is used but these materials cannot withstand [13]. Thus, rail beam is used due to its resistant to rust, workability, and durability. After the materials of the testbed are selected, the dimension of the testbed needs to be determined. A large testbed is very difficult to handle and requires a lot of space and may raise safety concern, where big testbed can cause injuries if it falls from designated area. The position of the test bed is also very important; most of the experiment done by other researcher are in vertical position and they are using joule heating element [10, 15, 16]. Due to different uses of heating elements, Peltier module, and fan as cooling element, which both has flat surfaces, the position decided in this testbed will be in horizontal position. The testbed needs to be put on a flat and stable surface to ensure that data collected is reliable and good to be compared with another researcher. Figure 1a shows the experiment setup rig, while Fig. 1b shows the arrangement of the SCP with Peltier module.

Electrical circuits are being used in this experiment setup are Peltier circuit, load cell circuit, and temperature sensor circuit. Peltier circuit for this experiment consists of Peltier module, Wheatstone-bridge and power supply. Peltier module used in this experiment is TEC-1-12715, with power rating of 130 W. Temperature sensors that are selected in this experiment are either PT100 or PT1000 from LABFACILITY. This low-cost, surface mount device provides an accurate reading and have wide range of temperature from −50 to 1000 °C. The data were recorded using Agilent Data Acquisition devices or DAQ that are connected to a regular lab computer, as shown in Fig. 2.

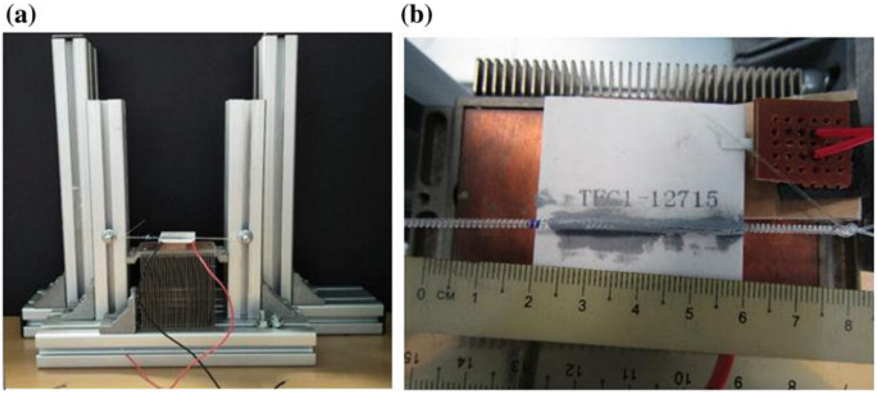


Fig. 1 **a** Experiment setup. **b** Peltier module heating the SCP that are coated with thermal paste during the experiment

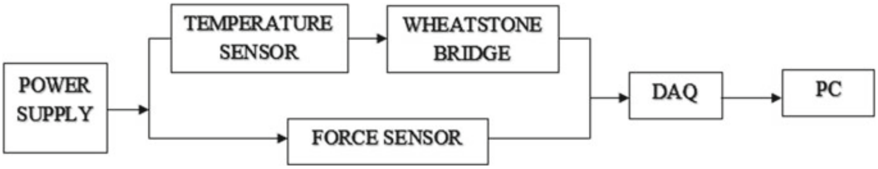


Fig. 2 Block Diagram of circuit

2.2 Material/Sample Preparation

The material used in this research is Berkeley Trilene Big Game nylon fishing line. The nylon fishing lines were cut with 60 cm length and a loop were made at both ends of the fishing lines. To make the artificial muscles or Super-coiled polymer (SCP), the nylon fishing lines were twisted using a portable drill or DC motor until it forms a helix shape or a coil. It is important to ensure the strings are tensed, so that the strings are easily twined when the portable drill is turn on. This method was also used by another researcher in [17]. It can be easily used in low-cost 3D printed prosthetic that are made for small amputees that needed changes every year.

2.3 Conduct of Experiment

Before the SCP were put on the test bed, it will undergo the annealing process. Annealing process is a heat treatment process, which alters the microstructure of a material to change its mechanical or electrical properties [18–20]. These tests were done using KHIND kitchen oven, where the dial on the oven was set at 100 °C in four different combinations of annealing process. The combinations are:

(a) Fast Heating, Fast cooling (FHFC)—Case 1
(b) Fast Heating, Slow Cooling (FHSC)—Case 2
(c) Slow Heating, Slow Cooling (SHSC)—Case 3
(d) Slow Heating, Fast Cooling (SHFC)—Case 4.

The oven was preheated at 100 °C for 3 min before the experiment was conducted. Each SCP actuator had a loop made at both ends for easy installation on the test bed of the annealing process, as shown in Fig. 3a. To differentiate fast heating and slow heating, the heating element of the kitchen oven was changed into both sided for fast heating, and one sided for slow heating.

The steel bar was then placed in the middle of regular kitchen oven to ensure the heating temperature of the string was distributed evenly during the annealing process. The heating process was done for 3 min, while the cooling process was done for 4 min each, for each combination and test specimen. The cooling process was determined as slow cooling when the door of the oven was closed. and as fast

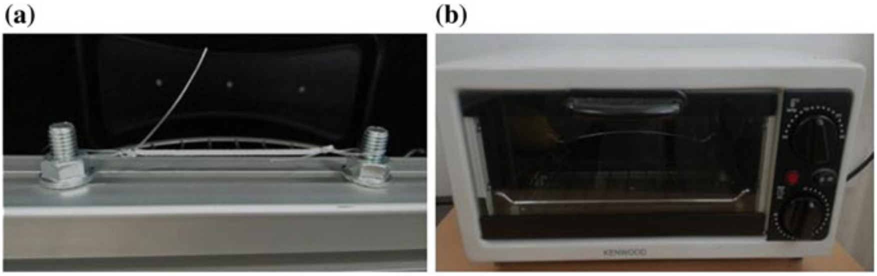


Fig. 3 **a** SCP during annealing process. **b** Oven used in the experiment

cooling the oven door was open. After the annealing process, the string was attached across a Peltier module that was attached to another apparatus as shown in Fig. 1b.

This experiment was conducted to test the performance of the nylon-based artificial muscles that had undergone the annealing process. A thermal paste and a cooling fan were constructed in the test bed, to ensure evenly distribution of temperature during the test. A temperature sensor was attached on one side of Peltier module to measure the temperature on Peltier module. In this testing, we used RTD sensor Class B from Labfacility. Each combination had two sets of SCPs that had undergone the annealing process and each string were tested for 5 times. The data of these experiments were collected using Agilent DAQ devices that were connected to a PC. The types of data collected during this experiment were the temperature data in Celsius and tensile force data in Newton from the load cell.

3 Result and Discussion

3.1 Durability

The life span or durability for the SCP is determined by the number of cycles that the SCP can expand and contract before unravelling. Once the SCP unravels, it is unable to perform reliably. Figure 4 shows the performance of SCP in five cycles. This SCP undergoes Fast heating fast cooling (FHFC) annealing process. Comparing four different cases of annealing process, FHFC gave the best result as it only unraveled after 4 cycles, while the others unraveled after only 1 or 2 cycles, as shown in Table 1.

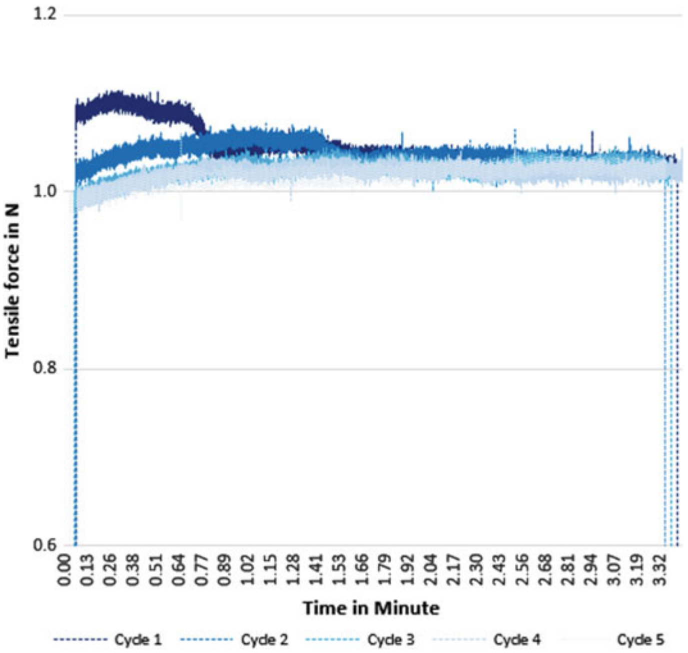


Fig. 4 Tensile force produced by the SCP that undergone FHFC conditions for 5 cycle

Table 1 Durability for different cases

Case	No. of cycle before unraveled
1	4
2	2
3	1
4	2

3.2 Amount of Tensile Force

The second part is to determine the maximum tensile force that the SCP can exert for each of the four cases. Ideally, the higher tensile force would be preferable as this would indicate that the SCP can pull or carry heavier loads. Figure 5 shows the performance graph for different SCPs that have undergone different annealing processes. The SCP in the first case gave the highest maximum tensile force at about 1.12 N. The other SCPs are not far behind, producing between 1 and 1.05 N of tensile force (Table 2).

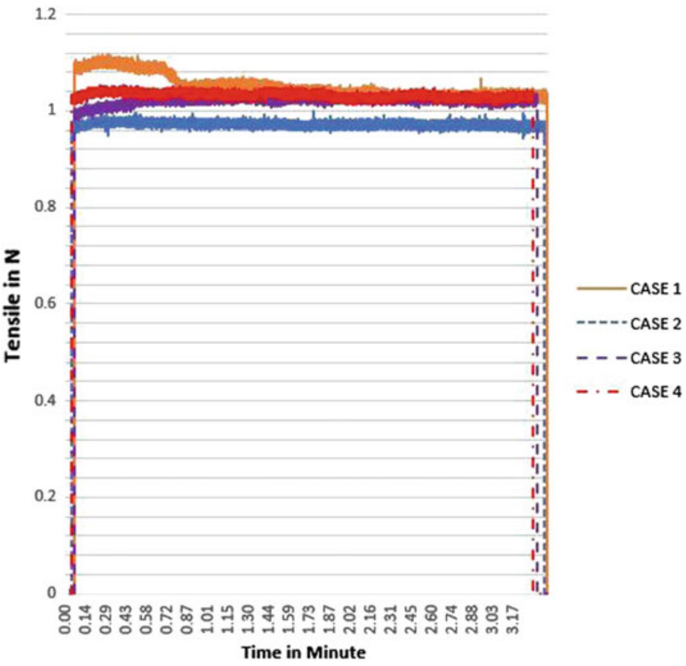


Fig. 5 Maximum tensile force for 4 different annealing cases

Table 2 Maximum tensile force exerted by the SCP for each case

Case	Max. tensile force (N)
1 (FHFC)	1.116211
2 (FHSC)	1.001404
3 (SHFC)	1.046448
4 (SHSC)	1.054138

3.3 Time Taken to Reach Maximum Force

The third part is looking into the amount of time taken for the SCP to exert the maximum tensile force. Generally, each SCP shows similar pattern where the tensile force increases from zero to maximum in a relatively short amount of time. However, it is noticed that certain cases, take slightly more time compare to others. As shown in Fig. 6 and Table 3, the SCPs in case 2 and 3 were the fastest, for which it only took 0.05 min to reach to the maximum point. SCPs in case 1 and 4 took a slightly longer time at 0.07 min.

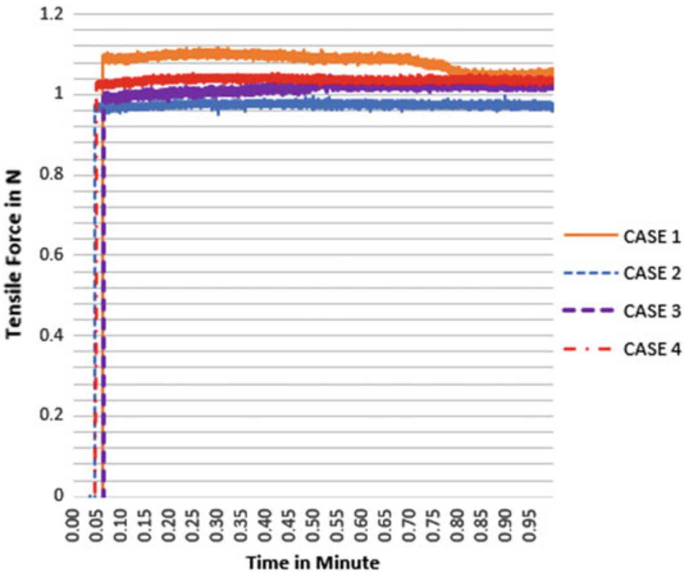


Fig. 6 Graph showing time taken to reach maximum tensile force for 4 different cases

Table 3 The time taken to reach maximum tensile force for each case

Case	Time (min)
1 (FHFC)	0.07
2 (FHSC)	0.05
3 (SHFC)	0.07
4 (SHSC)	0.05

4 Conclusion

This paper describes the work done to study the impact of different annealing processes on the life span and performance of nylon-based strings, also known as Super-coiled polymer (SCP). The SCPs were prepared using four different annealing processes based on the rate of heating and cooling. The results from the conducted experiments show that the SCP that was prepared using fast heating fast cooling (FHFC) annealing process, has the longest life span or the durability as well as being able to exert highest tensile force compared to other cases. However, the FHFC SCP took slightly, longer time to reach the maximum tensile force.

Acknowledgements This research is funded under UNITEN Internal Research Grant (UNIIG J510050696). The authors would like to thank those who had directly and indirectly provided insights which greatly assisted this research.

References

1. Maziz A, Concas A, Khaldi A, Stålhand J, Persson NK, Jager EWH (2017) Knitting and weaving artificial muscles. Sci Adv 3(1):1–12
2. Aach M et al (2014) Voluntary driven exoskeleton as a new tool for rehabilitation in chronic spinal cord injury: a pilot study. Spine J 14(12):2847–2853
3. Au SK et al (2015) System identification of human joint dynamics. IEEE Int Conf Rehabil Robot 18(1):55–87
4. Klute GK, Czerniecki JM, Hannaford B (1999) McKibben artificial muscles: pneumatic actuators with biomechanical\nintelligence. In: 1999 IEEE/ASME international conference on advanced intelligent mechatronics (Cat. No. 99TH8399), pp 1–6
5. Tondu B, Lopez P (2000) Modeling and control of McKibben artificial muscle robot actuators. IEEE Control Syst Mag 20(2):15–38
6. Pei Q, IEEE (2009) Artificial muscles based on synthetic dielectric elastomers. In: 2009 annual international conference of the ieee engineering in medicine and biology society, vols 1–20, pp 6826–6829
7. Zheng W et al (2011) Artificial muscles based on polypyrrole/carbon nanotube laminates. Adv Mater 2011(23):2966–2970
8. Meng Q, Hu J (2009) A review of shape memory polymer composites and blends. Compos. Part A Appl Sci Manuf 40(11):1661–1672
9. Haines CS, Li N, Spinks GM, Aliev AE, Di J, Baughman RH (2016) New twist on artificial muscles. Proc Natl Acad Sci U S A 113(42):706–728
10. Haines CS et al (2014) Artificial muscles from fishing line and sewing thread. Science 343 (6173):868–872
11. Mirvakili SM et al (2014) Simple and strong: twisted silver painted nylon artificial muscle actuated by Joule heating. Proc SPIE 9056(905601):1–10
12. Saharan L, Sharma A, Jung de Andrade, M, Baughman RH, Tadesse Y (2017) Design of a 3D printed lightweight orthotic device based on twisted and coiled polymer muscle: iGrab hand orthosis. Proc SPIE 10164:1016428-1
13. Atikah NA, Leong YW, Adzly A, Chau CF, Salleh MSK, Izham ZA (2017) Control of non-linear actuator of artificial muscles for the use in low-cost robotics prosthetics limbs. In: IOP conference series: materials science and engineering, Paper no. 257
14. Cherubini A, Moretti G, Vertechy R, Fontana M (2015) Experimental characterization of thermally-activated artificial muscles based on coiled nylon fishing lines. AIP Adv 5 (067158):1–11
15. Yin H, Zhou J, Li J, Joseph VS (2018) Fabrication and properties of composite artificial muscles based on nylon and a shape memory alloy. J Mater Eng Perform 27(7):3581–3589
16. Netland Ø et al (2017) On the control and properties of supercoiled polymer artificial muscles. Sens Actuators A Phys 2(3):1–6
17. Haines CS (2014) How to make an artificial muscle out of fishing line. Sci Friday, pp 1–6
18. Lynch-Aird N, Woodhouse J (2017) Annealing process of nylon string. Mater (Basel) 10 (497):4
19. Babatope B, Isaac D (1992) Annealing of isotropic nylon-6, 6. Polymer (Guildf) 33(Dec 1990):1664–1668
20. Zhang Q, Mo Z, Liu S, Zhang H (2000) Influence of annealing on structure of nylon 11. Macromolecules 33(16):5999–6005

Nonlinear Proportional Integral (NPI) Double Hyperbolic Controller for Pneumatic Actuator System

S. Jamian, S. N. S. Salim, S. C. K. Junoh, M. N. Kamarudin and L. Abdullah

Abstract In this paper, an evaluation of Nonlinear Proportional Integral (NPI) Double Hyperbolic controller, which is tested on a pneumatic actuator system is performed. The NPI Double Hyperbolic controller consists of two nonlinear functions that are added before proportional and integral gains. These functions are named as nonlinear proportional and nonlinear integral, in which they are embedded with a hyperbolic function. This controller scheme is designed via MATLAB/Simulink software and it is run four times for simulation and experimental work. The scheme of this controller includes the system's mathematical model, which is obtained by estimating model using System Identification Toolbox in MATLAB. In addition, the simulation and experimental works are tested using the step input. The performance of the proposed technique is compared between PI and NPI Double Hyperbolic controller based on the rise time and overshoot. The results show that the NPI Double Hyperbolic has better transient performance with smaller rise time and smaller overshoot via experimental work compared to the conventional PI controller.

Keywords PI controller · NPI double hyperbolic controller · Pneumatic actuator system · Transient response

S. Jamian (✉) · M. N. Kamarudin
Faculty of Electrical Engineering, Universiti Teknikal Malaysia Melaka,
Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia
e-mail: syamizajamian@gmail.com

S. N. S. Salim
Faculty of Electrical and Electronic Engineering Technology,
Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya,
76100 Durian Tunggal, Melaka, Malaysia

S. C. K. Junoh · L. Abdullah
Faculty of Manufacturing Engineering, Universiti Teknikal Malaysia Melaka,
Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia

Z. Zakaria and R. Ahmad (eds.), *Advances in Electronics Engineering*,
Lecture Notes in Electrical Engineering 619,
https://doi.org/10.1007/978-981-15-1289-6_21

1 Introduction

Pneumatic actuators have the advantages of a very low cost of implementation, clean, easy to work with [1], and because of that factors, the used of pneumatic actuators grow vigorously widely in the industry. It was implemented in tasks like gripping, clamping, and spraying [2]. Unfortunately, it has the advantages such as the nonlinearities, affected by the high compressibility of the air, oscillation of the air pressure, expensive, and heavy for high power application [3], which cause the difficulties when it comes to use the pneumatic actuator system.

Research on pneumatic position control has grown slightly since 20 years ago such as in the study conducted by [4–6]. Based on previous studies, most of the researchers stated that the dynamic model of the pneumatic actuator system is indicated by particular nonlinearities. Hence, when the conventional was used, it was not able to give a good performance. Then for the past year, many researchers came up with the new proposed control strategies, or made an invention to the standard controller, such as in the study by [7], a modification of PID by computing the friction compensated, the bounded integral action and position feed forward. Next, in [8], the cascade PID controller for a practical pneumatic system with good disturbance rejection was introduced. Author of [9] proposed a control technique of fuzzy proportional integral derivative which was accompanied with an asymmetric fuzzy compensator.

Nonlinear controllers were also implemented to improve the performance of the controller. Many studies use nonlinear controller to modify the conventional PID. The nonlinear PID can accomplish both great static and dynamic performances because the changes in characteristics of the nonlinear function, along with the ideal changes, process the parameters. The application of NPID in pneumatic actuator is stated in the studies by [10–12]. As stated previously, this paper used nonlinear PI Double Hyperbolic to analyse the performance of the controller in pneumatic actuator system. This control algorithm was implemented previously by researcher [13], in it was applied on the XY table ball-screw drive system.

This paper is organized as follows: A modeling of a pneumatic actuator using system identification toolbox in MATLAB, is presented followed by the experimental set up. Next, the description of the conventional PI controller and the Nonlinear PI Double Hyperbolic controller is shown in the next section. Lastly, the results of the simulation and the experiment are evaluated.

2 Methodology

2.1 Experimental Set up

The system consists of

(i) 5/3 bidirectional proportional control valve (Enfield LS-V15 s),
(ii) double acting with single rod cylinder (model ACTB-200-S01200),
(iii) data acquisition card (NI-PCI-6221 37-pin card),
(iv) pressure sensor (GEMS),
(v) position sensor,
(vi) personal computer,
(vii) open LabBox,
(viii) PANTHER air compressor.

The pneumatic cylinder with 300 mm stroke for both sides is fixed on the base. One side of the cylinder bar is associated with the carriage and drives an inertial load on the rails. The data acquisition and control schedules executed under real-time workshop in MATLAB are acknowledged by a PC furnished with a PCM card. Two pressure sensors are placed on the two sides to quantify the different pressures between two chambers. Figure 1 shows the experimental set up, which was used to validate the proposed method.

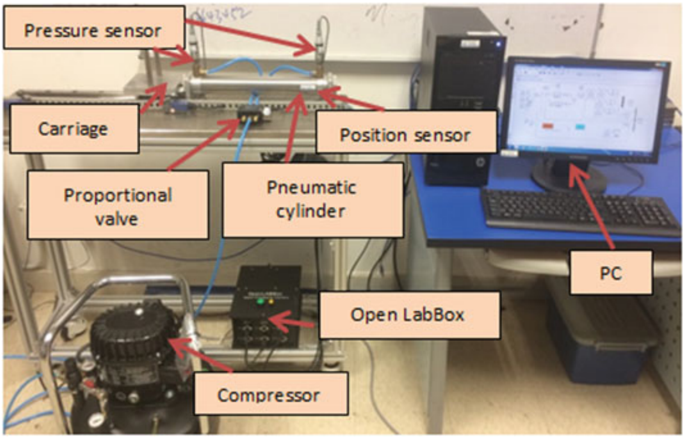


Fig. 1 A figure of pneumatic actuator system experimental set up

2.2 System Model

System identification is applied to obtain the transfer function of the system. The input and output data were collected for the first stage based on the open loop system with sampling time of 0.01 s. The data is exported to the workspace. The input signals with multi amplitude and frequency sine wave were used, where 4000 number of data was collected. This study chose transfer function model as a model structure of the system. The transfer function generated is represented as follow:

$$Y(s) = G(s)U(s) + E(s) \tag{1}$$

where,

$$G(s) = \frac{NUM(s)}{DEN(s)} \tag{2}$$

G(s) is the desired transfer function relating the input u(t) to the output y(t). Transfer function of the system comes in 3 poles and 2 zeroes. Third order system was used in this study as in the previous study. Most of the researchers used third and fourth order system, so it is easy to be cooperates with the complex system. The estimation of the parameters is computed by iterative search for the model, through the Trust-Region Reflective Newton (Isqnonlin) method with 50 numbers of iterations. Equation (3) shows a continuous transfer function.

$$G(s) = \frac{1.478s^2 + 1.122s + 0.05463}{s^3 + 0.7467s^2 + 0.1132s + 0.06718} \tag{3}$$

3 Controller Design

3.1 Proportional Integral (PI) Controller

In this section, the PI controller is designed via MATLAB/Simulink software to perform the simulation and experimental work. The control scheme of PI controller

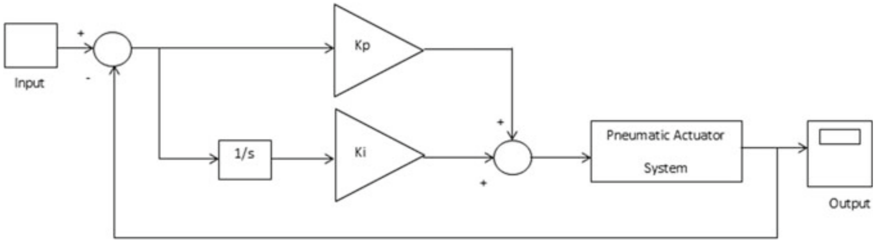


Fig. 2 Block diagram of PI controller

is shown in Fig. 2. Basically, the PI controller consists of two gains; proportional and integral gains. In order to obtain a suitable PI parameter, an experimental work is conducted. Firstly, the proportional gain is tuned, while the integral and derivative are set to zero. After that, the integral gain is tuned. Meanwhile, the derivative gain is set to zero due to the instability system. The PI parameter tuning method is shown in Table 1 and a schematic flowchart is shown in Fig. 3.

Table 1 Parameters of the PI controller

Control strategy	Control parameter		
	Name of parameters	Abbreviation	Value
PI	Proportional gain	K_p	9.3144
	Integral gain	K_i	3.6541
	Derivative gain	K_d	0

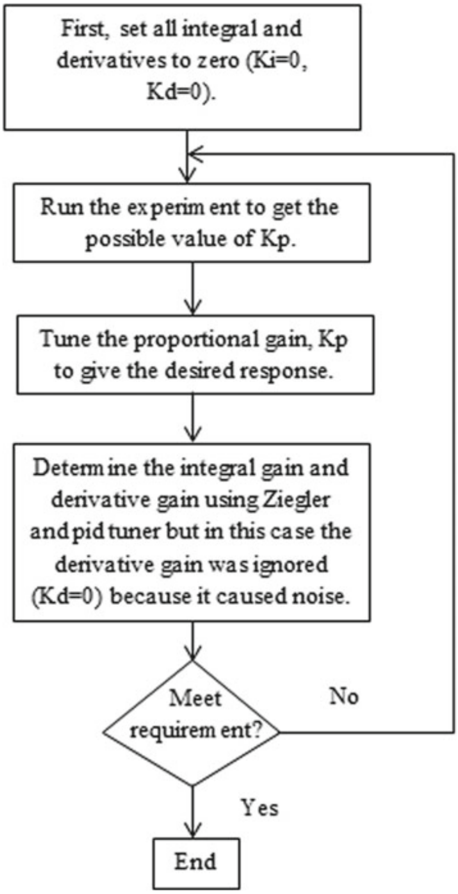

First, set all integral and derivatives to zero (Ki=0, Kd=0).
Run the experim ent to get the possible value of Kp.
Tune the proportional gain, Kp to give the desired response.
Determ ine the integral gain and derivative gain using Ziegler and pid tuner but in this ca se the derivative gain wa s ignored (Kd=0) because it caused noise.
Meet requirem ent?
No
Yes
End

Fig. 3 PI controller design procedure

3.2 Nonlinear Proportional Integral (NPI) Double Hyperbolic Controller

Next, the NPI Double Hyperbolic controller is designed using MATLAB/Simulink software as shown in Fig. 4, which consists of a control scheme with two nonlinear: nonlinear proportional (Np) and nonlinear integral (Ni) that are added before proportional and integral gains, respectively. The Np and Ni algorithms are embedded with hyperbolic function as stated in Eqs. (4) and (5), in which these nonlinear functions were already used and tested on another application by previous researcher [13]. Due to the effectiveness of this controller on machine tool application, this study is inspired to test this controller on pneumatic actuator application. Thus, the Np is designed by increasing the gain when the error is increased and vice versa. On the contrary, the Ni is designed by decreasing the gain when the error is increased. It means that the Np and Ni are able to correct the PID controller based on the value of the error. The parameters of the Np and Ni for this study are tabulated in Table 2. Furthermore, this controller is tested via simulation and experimental work for four times with different parameters of Np and Ni. These parameters are tuned and selected based on the performance of the system. The control signal of NPI Double Hyperbolic controller is derived in Eq. (6).

$$\text{Nonlinear proportional}, N_P = 1 + f \cdot \left(1 - \sec\, h\left(g \cdot e_p\right)\right) \tag{4}$$

$$\text{Nonlinear integral}, N_I = \frac{1}{p + (q \cdot (1 - \sec\, h(r \cdot e_I)))} \tag{5}$$

Control signal of NPI Double Hyperbolic,

$$U_{DOUBLE} = K_P(N_P.\, e(t)) + K_I\left(N_I \int_0^t e(t)dt\right) \tag{6}$$

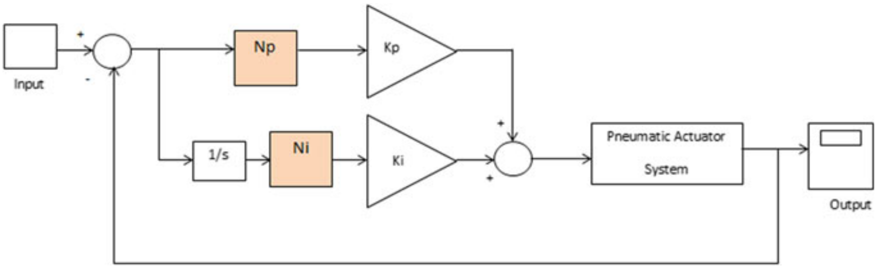


Fig. 4 Block Diagram of NPI Double Hyperbolic

Table 2 Parameters determination of Np and Ni

	f	g	e_p	p	q	r	e_i
P1	2.0	0.6	0.8	0.3	0.04	0.07	0.7
P2	5.0	0.6	0.8	0.3	0.04	0.07	0.7
P3	20.0	0.6	0.8	0.3	0.04	0.07	0.7
P4	20.0	0.6	5.0	0.3	0.04	0.07	0.7

4 Result and Discussion

In this section, the results for simulation and experimental work are conducted via MATLAB/Simulink software using a step signal with amplitude 7. The results are evaluated based on transient response (rise time and overshoot). Basically, a better transient response shows smaller rise time and smaller overshoot. The results for both controllers are shown in Fig. 5 (simulation) and Fig. 6 (experimental), and the results are tabulated in Table 3. For simulation, the rise time of NPI Double Hyperbolic is 0.5268 s, which is smaller than PI controller with rise time of 0.5294 s. The results for experimental work also show better transient with smaller rise time when NPI Double is applied compared to PI controller. Furthermore, the overshoot during simulation shows that the NPI Double 1 has a higher overshoot value compared to the other NPI Double and PI controllers. However, during experimental work, the NPI Double 1, 2, 3, and 4 show smaller overshoot compared to PI controller. It means that the nonlinear function in the NPI Double Hyperbolic controller is able to cater the nonlinearity of the pneumatic application. Based on [13], the nonlinear proportional is able to produce higher gain when the error is high and vice versa, while the nonlinear integral produces smaller gain when the error is high and vice versa.

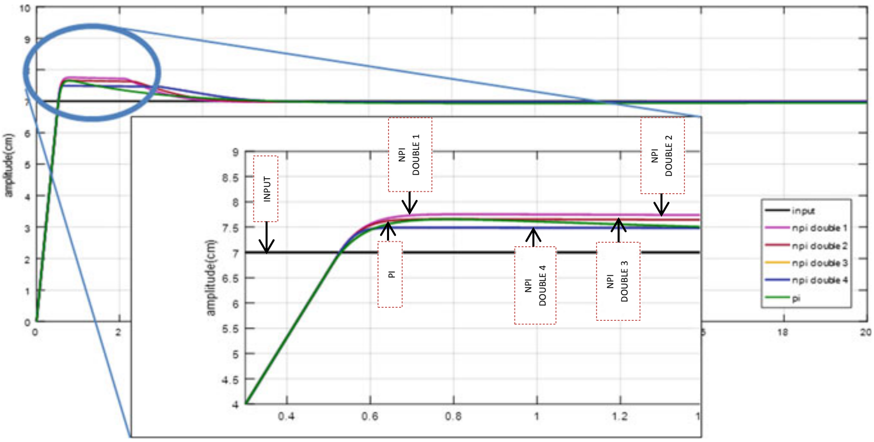


Fig. 5 Simulation result of PI and NPI double hyperbolic

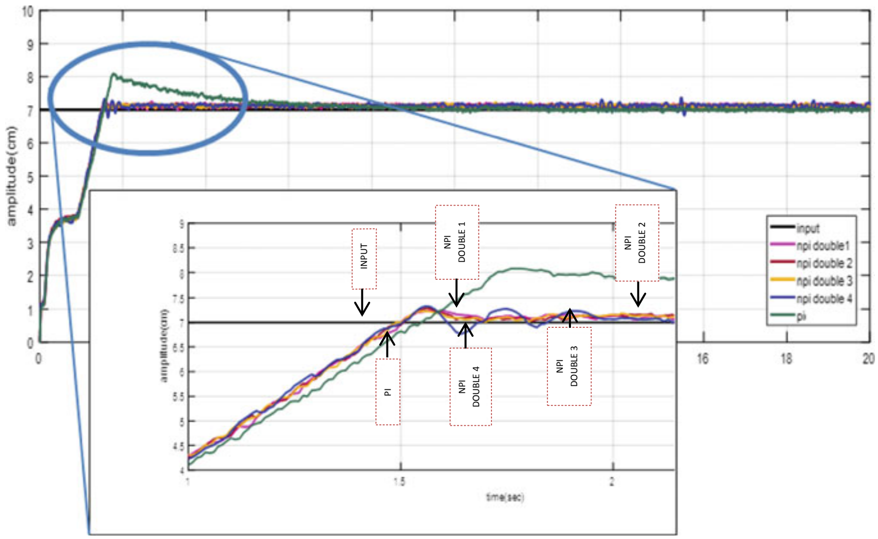


Fig. 6 Experimental result of PI and NPI Double Hyperbolic

Table 3 Simulation and experimental result of PI and NPI Double Hyperbolic

	Simulation		Experimental	
	Rise time, Tr (s)	Overshoot (%)	Rise time, Tr (s)	Overshoot (%)
PI	0.5294	9.3	1.5642	15.4
NPI Double 1	0.5268	10.7	1.5010	3.4
NPI Double 2	0.5268	9.4	1.5065	4.1
NPI Double 3	0.5268	7.0	1.4891	3.3
NPI Double 4	0.5268	7.0	1.5039	4.7

5 Conclusion

This paper evaluates the two different controllers: PI controller and NPI Double Hyperbolic controller for pneumatic actuator system. The controllers are designed via MATLAB/Simulink software including a controller scheme. The NPI Double Hyperbolic controller shows better transient performance in terms of rise time and overshoot compared to PI controller. The NPI Double Hyperbolic at P1, P2, P3, and P4 shows rise times 1.5010, 1.5065, 1.4891, and 1.5039 s, respectively, and overshoots 3.4%, 4.1%, 3.3%, and 4.7%, respectively. Meanwhile, the PI controller shows rise time at 1.5642 s and overshoot of 15.4%. The NPI Double Hyperbolic has proven that the controller produces excellent transient response in pneumatic system. It means that the hyperbolic algorithm in the NPI Double Hyperbolic controller is able to be used for different applications. The future study should include a stability analysis via either Popov plot or Lyapunov stability in order to obtain a suitable Np and Ni parameters.

Acknowledgements This research is supported by Universiti Teknikal Malaysia Melaka (UTeM) through Short Term-High Impact grant scheme (PJP-HI) with reference number of PJP/2017/FTK/HI/S01535. Authors are grateful to UTeM for supporting the present work.

References

1. Meng D, Tao G, Chen J, Ban W (2011) Modeling of a pneumatic system for high-accuracy position control. In: Proceedings of the 2011 international conference on fluid power mechatronics, FPM 2011, pp 505–510
2. Salim SNS, Rahmat MF, Faudzi AAM, Ismail ZH, Sunar NH, Samsudin SA (2014) Robust control strategy for pneumatic drive system via enhanced nonlinear PID controller. Int J Electr Comput Eng 4(5)
3. Hassan MY, Kothapalli G (2010) Comparison between neural network based PI and PID controllers, 6pp
4. Wang J, Pu J, Moore P (1999) A practical control strategy for servo-pneumatic actuator systems. Control Eng Pract 7(12):1483–1488
5. Richer E, Hurmuzlu Y (2000) A high performance pneumatic force actuator system part 2—nonlinear controller design. ASME J Dyn Syst Meas Control 122(3):426–434
6. Bone GM, Ning S (2007) Experimental comparison of position tracking control algorithms for pneumatic cylinder actuators. IEEE/ASME Trans Mechatron 12(5):557–561
7. Van Varseveld RB, Bone GM (1997) Accurate position control of a pneumatic actuator using on/off solenoid valves. IEEE Trans Mechatron 2(3):195–204
8. Saleem A, Taha B, Tutunji T, Al-Qaisia A (2015) Identification and cascade control of servo-pneumatic system using Particle Swarm Optimization. Simul Model Pract Theory 52:164–179
9. Yang G, Du J-M, Fu X-Y, Li B-R (2017) Asymmetric fuzzy control of a positive and negative pneumatic pressure servo system. Chin J Mech Eng 30(6):1438–1446
10. Acarman T, Hatipoglu C, Ozguner U (2001, June) A robust nonlinear controller design for a pneumatic actuator. In Proceedings of the 2001 American Control Conference (Cat. No. 01CH37148), Vol 6. IEEE, pp 4490–4495
11. Syed Salim SN, Rahmat MFA, Mohd Faudzi AA, Ismail ZH, Sunar N (2014) Position control of pneumatic actuator using self-regulation nonlinear PID. Math Probl Eng
12. Thanh TDC, Ahn KK (2006) Nonlinear PID control to improve the control performance of the pneumatic artificial muscle manipulator using neural network. Mechatronics 16(9):577–587
13. Junoh SCK, Abdullah L, Jamaludin Z, Anang NA, Chiew TH, Salim SNS, Retas Z (2017) Evaluation of tracking performance of NPID double hyperbolic controller design for XY table ball-screw drive system. In: 2017 11th Asian control conference (ASCC). IEEE, pp 665–670

CFOA Based Negative Floating Capacitance Multiplier

Shashwat Singh, Neeta Pandey and Rajeshwari Pandey

Abstract A floating negative capacitance multiplier circuit is proposed in the paper utilizing only two Current Feedback Operational Amplifiers (CFOAs) and total of three passive elements including two resistors for tuning the multiplication factor and a low valued integrated capacitor. The design does not need any matching conditions and possesses a very high multiplication factor with operational bandwidth of 1 MHz. The potential electronically tunable and temperature insensitive topology is also portrayed with no passive elements by incorporating Operational Trans-conductance Amplifier (OTA). Application in parasitic capacitance cancellation circuit is discussed in brief with SPICE simulations attached in support. A 100 Hz notch filter is also realized to observe the effectiveness of the design.

Keywords Current feedback operational amplifier (CFOA) · Electronically tunable · Floating negative capacitance multiplier · Operational trans-conductance amplifier (OTA) · Temperature insensitive

1 Introduction

Capacitor constitute an integral part of any electronic circuit or device as a whole. From filters to oscillators and gyrators to amplifiers, large valued capacitors are needed to achieve a low frequency region of operation. This large valued capacitor integration poses problem of large silicon area consumption in an IC. A rough estimate suggests that area occupied by 20pF capacitor is nearly equal to the silicon

S. Singh (✉) · N. Pandey · R. Pandey
Department of Electronics and Communication Engineering, Delhi Technological University,
Shahbad Daulatpur, Main Bawana Road, Delhi 110042, India
e-mail: singhsdtu@gmail.com

N. Pandey
e-mail: neetapandey@dce.ac.in

R. Pandey
e-mail: rajeshwaripandey@gmail.com

Z. Zakaria and R. Ahmad (eds.), *Advances in Electronics Engineering*,
Lecture Notes in Electrical Engineering 619,
https://doi.org/10.1007/978-981-15-1289-6_22

area covered by approximately thousands of transistors [1]. Thus, we need to devise a mechanism to obtain a large valued capacitance from a small one in the integrated circuit only. Here is when the capacitance multiplier circuits play its role.

Various capacitance multiplier circuits are available in literature comprising of both voltage mode and current mode circuits. But the voltage mode circuits suffer from the limitations like limited slew rate, lower gain bandwidth product and lower accuracy in comparison to the current mode circuits. Numerous circuits are already available in literature which employ varied analog active building blocks, such as Second Generation Current Conveyors (CCIIs) [2–6], Operational Trans Conductance Amplifiers (OTAs) and Operational Amplifier [7], Current-controlled Differential Difference Current Conveyors (CCDDCCs) [8], Voltage Differencing Buffered Amplifier (VDBA) [9], Current amplifier and Differential unity gain amplifier [10], Differential Voltage Current Conveyor (DVCC) [11–13], Current Differencing Transconductance Amplifier (CDTA) [14], Operational Transresistance Amplifier [15] and Current Feedback Operational Amplifier (CFOA) [16–18]. Some of the shortcomings of the previous work can be summarized as:

(a) large number of active blocks [6–8, 12, 13].
(b) more number of passive elements [6, 15].
(c) limited multiplication factor [1, 3, 8–10, 13–16].
(d) realized capacitance is grounded, which finds less usage in electronic circuits than floating ones [1–3, 6, 9, 14–17].

All the above reasons persuaded to develop a multiplier circuit, which overcomes these limitations to a great extent i.e. inculcate the good characteristics of all. CFOA block is adopted owing to its versatility and its advantages over the voltage amplifier blocks [19]. The simulations are performed using the commercially available AD844A macro model, which provides a compensating terminal giving more flexibility and helps in reduction of passive component count. Section 2 describes the recommended circuit that possesses only two CFOAs and two resistors with a source capacitance, which is a small valued capacitor. One of the potential realizations is also described having characteristics like electronic tunability and temperature insensitiveness. Sections 3 and 4 deal with the simulations with discussions and application part, respectively. Final concluding remarks are presented in Sect. 5.

2 Recommended Circuit

The CFOA design possesses an altogether different topology. A higher value of slew-rate is obtained owing to the use of current as the feedback signal. It ranges from 500 to 2500 V/ps as quoted by manufacturers of CFOAs, which is significant improvement from 1 to 100 V/ps of voltage mode circuits [20]. Thus, we can have

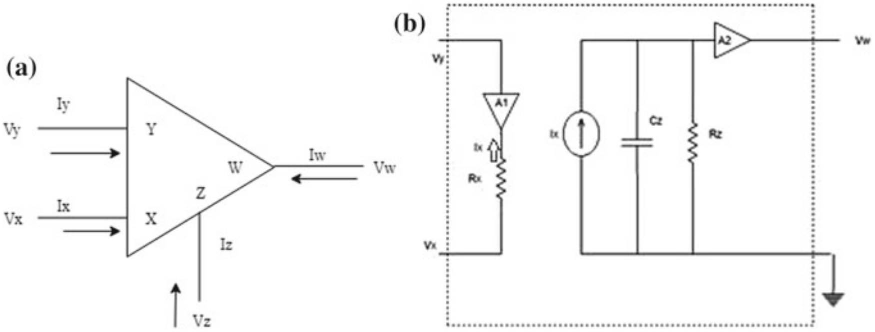


Fig. 1 **a** Symbolic diagram of CFOA **b** Equivalent circuit of CFOA [20]

a wide range of operational frequency range, which finds application in video signal conditioning. The symbol for the CFOA and its equivalent circuit is portrayed in Fig. 1. The port relationship, in matrix form can be written as:

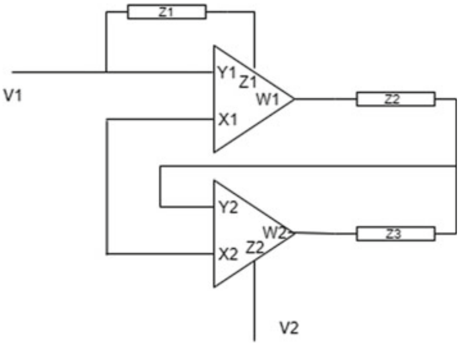
$$\begin{bmatrix} I_y \\ V_x \\ I_z \\ V_w \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & \pm 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} V_y \\ I_x \\ V_z \end{bmatrix} \tag{1}$$

The proposed circuit comprises of two CFOAs blocks and three passive elements including two multiplication factor tuning resistors and one source capacitor of small value, which can be integrated in IC. The recommended circuit is depicted in Fig. 2.

The input impedance of the circuit in Fig. 2 can be written as:

$$\frac{V_1 - V_2}{I_{in}} = -\frac{Z_1 Z_3}{Z_2} \tag{2}$$

Fig. 2 Proposed capacitance multiplier circuit



It can be interpreted from (2) that, we obtain a negative multiplied capacitance, where the multiplication factor can be determined in two possible ways:

A. When we take Z_1 as the source capacitor 'C_s', and Z_2 and Z_3 as resistors R_2 and R_3, we obtain:

$$Z_{in} = -\frac{R_3}{C_s R_2} \tag{3}$$

$$C_{eq} = -C_s \frac{R_2}{R_3}, \tag{4}$$

$$C_{eq} = -K_{m_1} C_s \tag{5}$$

$$K_{m_1} = \frac{R_2}{R_3} \tag{6}$$

Here, K_{m1} denotes the multiplication factor whose mathematical representation is given in (6).

B. When we take Z_3 as the source capacitor 'C_s', and Z_1 and Z_2 as resistors R_1 and R_2, we obtain:

$$Z_{in} = -\frac{R_1}{C_s R_2} \tag{7}$$

$$C_{eq} = -C_s \frac{R_2}{R_1}, \tag{8}$$

$$C_{eq} = -K_{m_2} C_s \tag{9}$$

$$K_{m_2} = \frac{R_2}{R_1} \tag{10}$$

Here, K_{m2} denotes the multiplication factor whose mathematical representation is given in (10).

So, we have two degrees of freedom for the provided capacitance multiplier circuit, which is a very attractive feature providing flexibility in designing electronic devices. The multiplication factor K_m, in general, is tunable with the variation in any one of the resistors or two in case of more precise tuning requirements. A modified second topology for the same structure is proposed, which is shown in Fig. 3, where Z_1 can be replaced by source capacitance. The CMOS schematic of the O.T.A. is depicted and its aspect ratio are mentioned in Fig. 4 and Table 1, respectively.

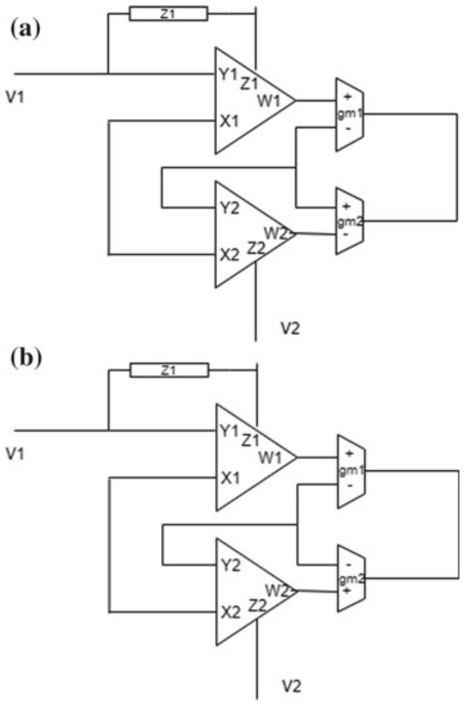


Fig. 3 **a** Positive capacitance multiplier circuit free of passive elements. **b** Negative capacitance multiplier circuit free of passive elements

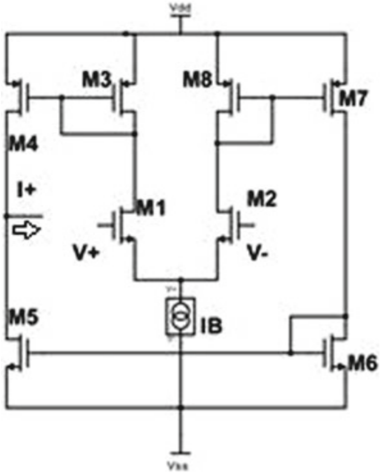


Fig. 4 CMOS structure for operational trans-conductance amplifier

Table 1 Aspect Ratio of O.T.A. in Fig. 4

Transistor	Aspect ratio (W/L)
M1, M2, M5, M6	18u/0.36u
M3, M4, M7, M8	36u/0.36u

We can conveniently repeat the same input impedance evaluation for Fig. 3a and b as we did earlier. For Fig. 3a, we have:

$$Z_{in} = \frac{Z_1 gm_1}{gm_2} \tag{11}$$

$$C_{eq} = C_s \frac{gm_2}{gm_1} \tag{12}$$

$$C_{eq} = C_s \sqrt{\frac{IB_2}{IB_1}} \tag{13}$$

where, gm_i represents the transconductance gain of the OTA and $gm_i \propto \sqrt{IB_i}$ in case of MOSFET in saturation region. And for Fig. 3b, we obtain:

$$Z_{in} = -\frac{Z_1 gm_1}{gm_2} \tag{14}$$

$$C_{eq} = -C_s \frac{gm_2}{gm_1} \tag{15}$$

$$C_{eq} = -C_s \sqrt{\frac{IB_2}{IB_1}} \tag{16}$$

Again, we have two degrees of freedom and hence, in total, there are 4 topologies for the modified circuit, out of which only two are represented for illustration purpose. Salient features of the modified circuit are:

(a) Passive elements are completely replaced in design.
(b) Both negative and positive multiplied capacitances can be simulated.
(c) The multiplied capacitance becomes electronically tunable with the help of bias current of the two OTAs.
(d) The multiplied capacitance is now temperature insensitive.

3 Simulations and Discussion

The functionality of the proposed circuit is verified through SPICE simulations using AD844A macro model of CFOA with ±5 V supply voltages. The O.T.A. is operated on supply voltage of ±1.8 V. All the topologies have canonic number of passive elements and do not require any passive component matching conditions. Simulations are performed for circuit in Fig. 3b.

The frequency response of the proposed circuit is depicted in Fig. 5. The bias current IB_1 fixed at 50 nA and IB_2 is varied from 50 nA and is tunable till several decades up to 250 mA. After this, the device parasitics dominate. The source capacitor has been taken as 1nF. By examining the magnitude and phase plot of the input impedance, we can understand that the variability of the multiplication factor is up to 5000. The transient response, which is attached in Fig. 6, demonstrates that the difference between phase of the input current and voltage is −90° i.e. voltage lead the current. Some deviations from ideal plot are observed, which are majorly due to internal parasitic elements, thus limiting the operational bandwidth. Therefore, the multiplier circuit has a practical frequency range from 100 Hz to 1 MHz.

Monte Carlo analysis is also performed using uniformly distribution randomly varied transport saturation current (I_s) and ideal maximum forward beta (BF) with a tolerance of 1%. Response is recorded and depicted in Fig. 7 in form of histogram. Detailed statistical parameters are listed in Table 2.

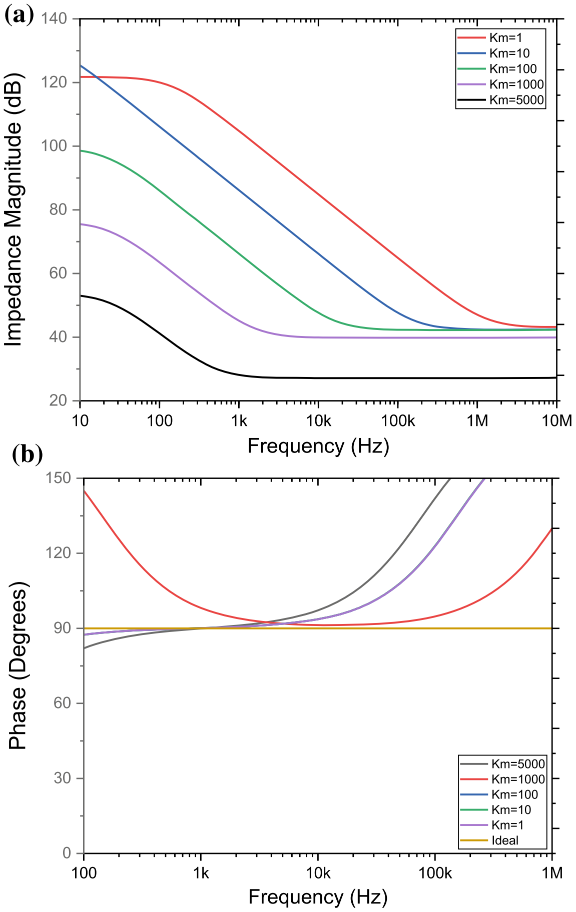


Fig. 5 **a** Magnitude plot of input impedance (in dB) for various multiplication factors. **b** Phase plot of input impedance (in degrees) for various multiplication factors

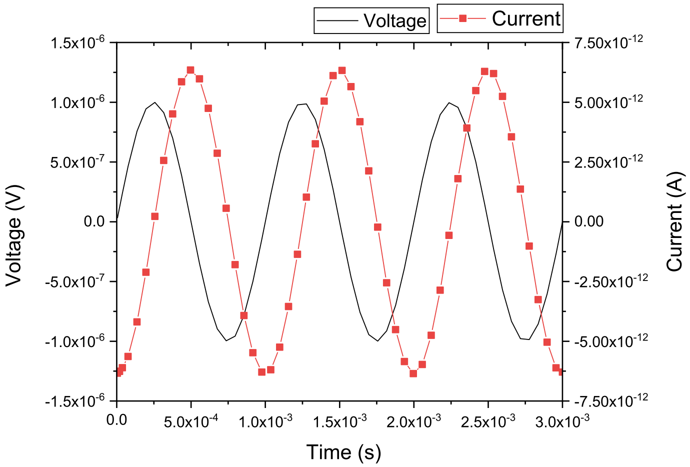


Fig. 6 Transient response depicting voltage leading by 90° w.r.t. current

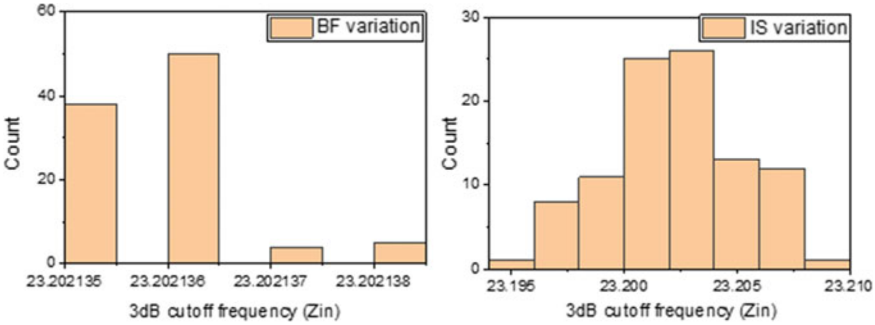


Fig. 7 Monte Carlo analysis

Any circuit's overall performance can be acceptable only after comparison of the circuit with the past literatures on various fronts. Table 3 compares those important performance parameters.

4 Application

The important objective of designing a negative capacitance multiplier can be observed in stray capacitance cancellation circuits [17], whose schematic is depicted in Fig. 8. The second terminal 'V_2' (Fig. 3b) is grounded and tuned to the value of the parasitic capacitance $C_{p.}$ Simulation results portray the significant

Table 2 Statistical parameters of monte-carlo analysis

Parameter varied (1%)	Mean	Median	Minimum	Maximum	Sigma
I_S	23.2023	23.2022	23.1946	23.2081	0.0029
BF	23.2021	23.2021	23.2021	23.2021	$8.5728*10^{-7}$

Table 3 Performance comparison

References	Nature of capacitance grounded (G)/ floating (F)	Active element utilized	No. of Active blocks	No. of passive element (s)	Multiplication factor
[1]	G	CCII and COA	2	2	100
[2]	G	CGCCII	1	0	NA
[3]	G	AD844	2	2	4
[6]	G	CCII	3	4	100,000
[7]	G and F both	OPAMP and OTA	3 and 5	0	1000
[8]	F	CCDDCC	3	0	1 to 10
[9]	G	VDBA	1	1	100
[10]	F	CA and DUA	2	0	300
[12]	F	DVCC, CCII and Digital Control Module	3	NA	NA
[13]	F	DVCC and CCCII	3	0	100
[14]	G	CDTA	1	2	~22
[15]	G	OTRA	1	3	160
[16]	G	CFOA	1	2	51
[17]	G	CFOA	2	2	NA
P.C.	*F*	*CFOA* and *OTA*	*4*	*0*	~*5000*

amount of cancellation achieved considering precision errors through Fig. 9. C_p is assumed to be 100nF for this demonstration purpose. It can be safely cancelled considering the multiplication range of our multiplier circuit. The negative simulated multiplied capacitance is represented by ‘$-C_c$’.

$$V_o = \frac{V_{in}}{1 + s\left(C_p - C_c\right)R_1} \tag{17}$$

$$Z_{in} = R_1 + \frac{1}{s\left(C_p - C_c\right)} \tag{18}$$

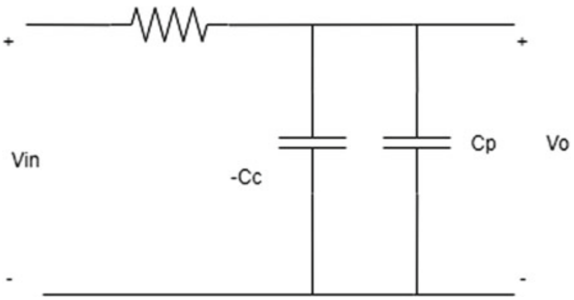


Fig. 8 Schematic of the parasitic capacitance cancellation circuit in low pass filter configuration

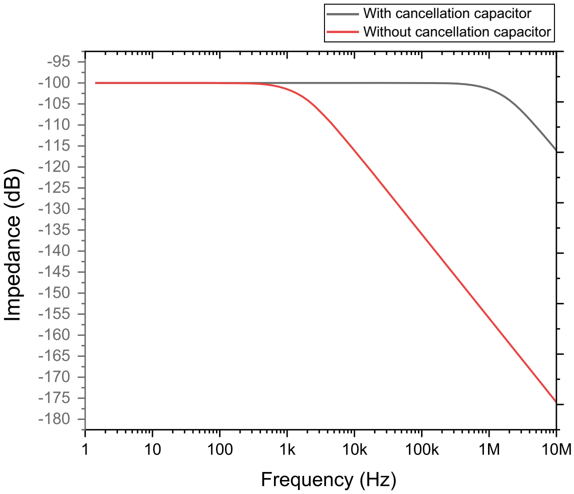


Fig. 9 Frequency response before and after employing capacitance cancellation circuit

The positive capacitance multiplier also finds major application in designing low cutoff frequency filters. The proposed active, low frequency and band reject filter using single CFOA is portrayed in Fig. 10a. The important point to note is that loading effect at the output i.e. at 'V_o' is completely eliminated by using CFOA. The complete circuit diagram is depicted in Fig. 10b. The CFOA1 (Y1, X1, Z1 and W1) and CFOA2 (Y2, X2, Z2 and W2) constitute the floating capacitor 'C1' and CFOA3 (Y3, X3, Z3 and W3) helps in obtaining the band rejection operation with no loading effect at the output. Leaving out the inductor, the complete circuit is perfectly integrable. Here, source capacitance 'C1' is used and positive capacitance multiplier is utilized, by which floating capacitor C_s is simulated and its mathematical relation can be referred to from (13). Table 4 lists the passive element values taken for simulation. Again, AD844A macro model and O.T.A. CMOS model of Fig. 4 have been used with ±5 V and ±1.8 V supply voltages, respectively.

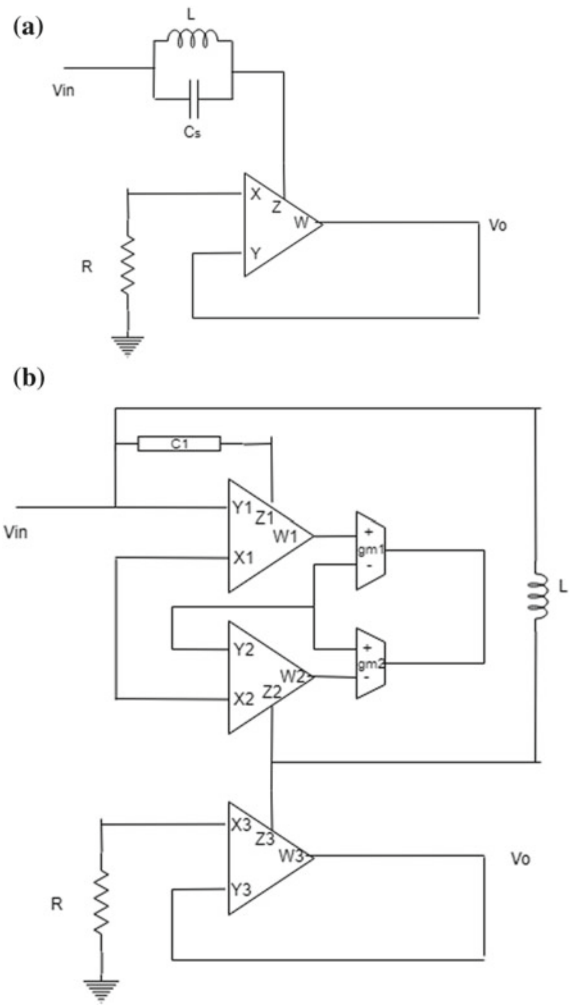


Fig. 10 **a** Single CFOA based equivalent filter design. **b** Proposed Notch filter with multiplier circuit deployed

Table 4 Component values

Component/centre frequency	f_o = 100 Hz
C	25 uF (5 nF * 5000)
L	100 mH
R	1 kΩ

The phase and magnitude responses are shown in Fig. 10. The transfer function H(s) can be written as follows:

$$H(s) = \frac{s^2 + \frac{1}{LC_s}}{s^2 + s\frac{1}{RC_s} + \frac{1}{LC_s}} \tag{19}$$

The central angular frequency is given by:

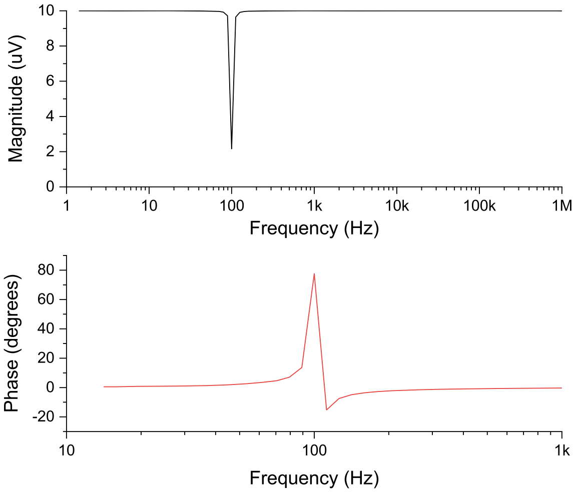
$$\omega_0^2 = \frac{1}{LC_s} \tag{20}$$

The quality factor (Q) and the bandwidth (B.W.) are calculated using the following relations:

$$Q = R\sqrt{\frac{C_s}{L}}, \quad B.W. = \frac{1}{2\Pi RC_s} \tag{21}$$

The rejection frequency has been adjusted to 100 Hz. The calculated and simulated value of the center frequency, from Fig. 11, are close to each other with error percentage of 0.65%. The quality factor obtained here is 15.8, which is much greater than unity. Bandwidth is found to be 6.36 Hz, which is well suited for a practical precise notch filter design.

Fig. 11 Band Reject filter magnitude and phase plot centered at 100 Hz



5 Conclusion

A negative floating capacitance multiplier circuit is proposed with attractive features of temperature insensitiveness, and electronic tenability. A very high multiplication factor of nearly 5000 is achieved using a resistorless design. No matching conditions are required and finds application in practical cases such as capacitance cancellation circuits. A single CFOA based low frequency notch filter using positive capacitance multiplier circuit is also presented with descent filter parameters. The SPICE simulations provide the soundness of the theoretical estimates.

References

1. Ferri G, Pennisi S (1998) A 1.5 V current-mode capacitance multiplier. In: Proceeding of the tenth international conference on microelectronics, pp 9–12. https://doi.org/10.1109/icm.1998.825555
2. Ferri G, Guerrini N (2001) High-valued passive element simulation using low-voltage low-power current conveyors for fully integrated applications. In: IEEE transactions on circuits and systems II: analog and digital signal processing, 48(4):405–409. https://doi.org/10.1109/82.933805
3. Khan AA, Bimal S, Dey KK, Roy SS (2002) Current conveyor based R- and C- multiplier circuits. Int. J. Electron. Commun 56(5):312–316. https://doi.org/10.1078/1434-8411-54100121
4. Minael S, Yuce E, Cicekoglu O (2006) A versatile active circuit for realising floating inductance, capacitance, FDNR and admittance converter. Analog Integrated Circ Sig. Process 47(2):199–202. https://doi.org/10.1007/s10470-006-4079-y
5. Yuce E (2006) Floating inductance, FDNR and capacitance simulation circuit employing only grounded passive elements. Int J Electron 93(10):679–688. https://doi.org/10.1109/TEL-NET.2017.8343536
6. Singh S, Jatin, Pandey N, Pandey R (2018) Precision capacitance multiplier with low power and high multiplication factor. In: 5th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, pp. 652–655. https://doi.org/10.1109/spin.2018.8474039
7. Ahmed MT, Khan IA, Minhaj N (1995) Novel electronically tunable C-multipliers. Electron Lett 31(1):9–11. https://doi.org/10.1049/el:19950018
8. Prommee P, Somdunyakanok M (2011) CMOS-based current-controlled DDCC and its applications to capacitance multiplier and universal filter. Int J Electron Commun 65(1):1–8. https://doi.org/10.1049/el:19950018
9. Unhavanich S, Onjan O, Tangsrirat W (2016) Tunable capacitance multiplier with a single voltage differencing buffered amplifier. In: Proceeding IMECS2016, Hong Kong
10. Al-absi MA (2017) New CMOS tunable floating capacitance multiplier. In: Int J Electron Lett 1–10. https://doi.org/10.1080/21681724.2017.1293167
11. Yuce E (2010) A novel floating simulation topology composed of only grounded passive components. Int J Electron 97:249–262. https://doi.org/10.1080/00207210903061907
12. Afzal N, Khan IA (2013) Digitally programmable floating impedance multiplier using DVCC. Int J Electron Commun Comput Technol 66(17):358–361
13. Siripruchyanan M, Jaikla W (2007) Floating capacitance multiplier using DVCC and CCCIIs. In: Proceedings of the International Symposium on Communications and Information Technologies (ISCIT ‘07), pp 218–221. https://doi.org/10.1109/iscit.2007.4392016

14. Biolek D, Vavra J, Keskin (2018) CDTA-based capacitance multipliers. In: A.Ü. Circuits system signal process, pp 1–16. https://doi.org/10.1007/s00034-018-0929-y
15. Singh S, Jatin, Pandey N, Pandey R (2019) Single OTRA based capacitance multiplier. Presented at international conference of advanced research and innovation, Delhi
16. Arslanalp R, Yücehan T (2007) Capacitance multiplier design by using CFOA. In: International symposium on communications and information technologies. https://doi.org/10.1109/siu.2015.7130102
17. Lahiri A, Gupta M (2011) Realization of grounded negative capacitance using CFOAs. Circuits Syst Signal Process 30:143–155. https://doi.org/10.1007/s00034-010-9215-3
18. Al-Absi MA, Abuelma'atti MT (2018) A novel tunable grounded positive and negative impedance multiplier. In: IEEE Transactions on circuits and systems II: express briefs. https://doi.org/10.1109/tcsii.2018.2874511
19. Palumbo G, Pennissi S (2001) Current feedback amplifiers versus voltage operational amplifiers. IEEE Trans Circuits Syst I 48(5):617–623. https://doi.org/10.1109/81.922465
20. Lidgey FJ, Hayatleh K (1997) Current-feedback operational amplifiers and applications. Electron Commun Eng J 9(4):176–182. https://doi.org/10.1049/ecej:19970404

Single Clock Diode Based Adiabatic Logic Family at Sub-90nm Regime

Sagar Jain and Neeta Pandey

Abstract The following manuscript investigates the performance of the proposed SC-DBAL (Single Clock Diode Based Adiabatic Logic) at 45 nm regime. The proposed logic belongs to the class of adiabatic circuits used for lower dissipation. The proposed logic uses a single phase split level triangular clock in its charging path and diode in its discharging path for reducing the rate of charging and discharging respectively. Simulations have been performed using standardized test bench for realistic simulations. Further, a 1 bit ALU using basic gates (NAND, NOR, XOR) has been realized using the proposed logic. The results show that the proposed logic dissipates least power among its closest competitor—DFAL while using a single clock and having reduced circuit complexity. TANNER EDA TOOL using PTM 45 nm HP (High—Performance) has been used for simulation. Power savings in range of 10–35% are observed over DFAL and 30–55% over conventional CMOS across frequencies ranging from 10 MHz to 200 MHz for different circuits.

Keywords Adiabatic logic · SC-DBAL · Logic gates · TANNER EDA tools

1 Introduction

CMOS technology proliferation has considerably improved the performance of the digital systems [1]. Reduction in feature size has further accelerated this process by increasing the number of transistors from 10 Million (350 nm) to 4 Billion (22 nm) on chip providing much more and better functionality than previous one [2]. However, this ever increasing number of transistors on chip, increases the power dissipated per chip. Active researches are being pursued in the VLSI domain to

S. Jain (✉) · N. Pandey
Department of Electronics and Communication Engineering, Delhi Technological University,
Shahbad Daulatpur, Main Bawana Road, Delhi 110042, India
e-mail: sagarjain1997@gmail.com

N. Pandey
e-mail: neetapandey@dce.ac.in

Z. Zakaria and R. Ahmad (eds.), *Advances in Electronics Engineering*,
Lecture Notes in Electrical Engineering 619,
https://doi.org/10.1007/978-981-15-1289-6_23

reduce the power dissipated on chip [3]. Reducing the feature size might be seen as plausible solution to this problem as lower supply is used, but the disproportionate reduction of threshold voltage and supply voltage leads to performance degradation due to lower current drive and rise in leakage current.

Power dissipated in digital logic circuits mainly consist of two primary components namely Static and Dynamic [3]. Static Power Dissipation occurs in CMOS when the supply and ground are directly connected for a brief interval during switching. It can be reduced and sometimes completely eliminated depending upon the structure; by changing the placement of inputs in the CMOS network. Dynamic Power Dissipation is the dominant component and governs the power dissipated in digital logics. It is the energy dissipated in charging the node of the circuit from the power supply. When the LOGIC changes, the charge on the node is discharged to ground. It is given by the following formula:

$$E_{CMOS} = \alpha C_L V_{DD}^2 \tag{1}$$

where, α is the switching factor, C_L is the load capacitance and V_{DD} is the supply voltage.

Adiabatic Logic provides an alternate implementation, wherein the charge taken from the supply for logic evaluation is recovered back by it after the evaluation [5]. It borrows the idea from the thermodynamic adiabatic operation, where no heat is taken or given to the surroundings under ideal conditions [6]. Unlike CMOS implementation, which uses constant power supplies, adiabatic logic uses time varying clocks (triangular, sinusoidal or trapezoidal) to recover the charge from the nodes after logic evaluation [7–11]. Reduced power dissipation accompanied with compatibility to the existing EDA tools has made adiabatic logic productive for design. Operating transistors at sub 90 nm regime introduces the problem of leakage in the transistors present on the chip. It is an unavoidable phenomenon leading to static losses in CMOS circuits, threshold voltage is increased to ensure low leakage leading to trade-off between power and performance. High threshold voltages, lower gate overdrive voltage (V_{GS}-V_{TH}), reduces the adiabatic charging fraction in adiabatic logic circuits, leading to higher adiabatic losses. A host of other problems such as Hot Carrier Junction (HCI) and Bias Temperature Instability (BTI) too creep up as the feature size is reduced [4].

Another challenge for adiabatic circuits is the requirements of power clocks. Existing single ended adiabatic families [12, 13] not only use more than one clock, but also more than one phase for their operation. This leads to increased complexity of circuitry for clock generation as well as the routing of the clocks on the chip [8]. Another major concern of these clocks is that, the generation circuitry should be able to maintain optimum phase relation between the clocks; else the gains expected might not be attained. GFCAL [7] proposed in the year 2008, uses a single clock in its operation, hence does not need to maintain the phase relationship, lessening the burden on the clock generation circuit. For reduced power dissipation, recently the use of split phase clocks as opposed to full swing clocks has garnered attention [11]. DFAL operating on split level clocks is proven to perform better than existing adiabatic logic families [7, 12, 13]. Although DFAL uses split level sinusoidal clocks for its operation, the authors do not provide any circuitry for generation of these clocks. Generating them from two different sources might lead to problem of phase inconsistency, reducing the expected gains. Researchers have also shown that triangular clocks dissipate least power among sinusoidal and trapezoidal clocks [7, 14]. Upadhyay et al. [14] operates on split level triangular clock for lower power dissipation, but requires 2 clocks leading to higher burden on clock generation.

Motivated by the work of previous researchers we decided to use single phase split level triangular clock, having least power dissipation and lower circuit complexity, and propose a new adiabatic logic family called Single Clock Diode Based Adiabatic Logic Family (SC-DBAL). Although generation of triangular clock is difficult, but because it is single, no phase consistency circuit is required. The rest of the paper is as follows: Section 2 delves into the adiabatic logic circuits and discusses about DFAL. Section 3 discusses about the operation of the proposed SC-DBAL. Section 4 discusses the design of 1 bit ALU using proposed logic by utilizing basic gates like NAND, NOR and XOR. Section 5 provides the simulative investigation of proposed, DFAL and CMOS Logic, using a simulation test bench given in [4]. Conclusion is given in Sect. 6.

2 Adiabatic Logic Circuits

2.1 Adiabatic Logic

Owing to the linear region of operation in adiabatic logic style, the charging operation can be thought of as an RC circuit operated by an increasing supply voltage, wherein C is the node capacitance. During discharging, the circuit can be modeled as an RC circuit having a charged capacitor operated by decreasing voltage supply. The energy dissipated in adiabatic logic can be given by the following:

$$E_{\text{DISSPATION}} = E_{\text{CHARGING}} + E_{\text{DISCHARGING}} \tag{2}$$

$$= I^2 R_{\text{CHARGING}} T + I^2 R_{\text{DISCHARGING}} T \tag{3}$$

$$= (R_{\text{CHARGING}} C_L + R_{\text{DISCHARGING}} C_L) C_L V_{DD}^2 / T \tag{4}$$

where, time period of the supply is given by ‘T’, charging/discharging path resistances are $R_{\text{CHARGING}}/R_{\text{DISCHARGING}}$ respectively, load capacitance is given C_L and Supply Voltage is denoted by V_{DD}. By (4) $E_{\text{DISSIPATION}}$ tends to zero when T >> (R_{CHARGING} + $R_{\text{DISCHARGING}}$)C_L. This is referred to as adiabatic loss in the literature.

$$E_{\text{NON-ADIABATIC}} \alpha\, C_L V_{THP}^2 \tag{5}$$

$$E_{\text{ADIABATIC}} \alpha (R_{\text{CHARGING}} C_L + R_{\text{DISCHARGING}} C_L) C_L V_{DD}^2 / T \tag{6}$$

Adiabatic logic circuits suffer from non-adiabatic losses given by (5), which is proportional to the threshold voltage and load capacitance. The devices turn OFF when the supply reaches below the threshold level leading to incomplete charge recovery ex. Quasi Adiabatic Logic Families (DFAL, etc.) [11]. Fully Adiabatic Logic Families like SCRL [12] completely eliminate such losses. Equations (5) and (6) give dependence of non-adiabatic losses and adiabatic losses on circuit and component parameters.

2.2 DFAL

DFAL stands for Diode Free Adiabatic Logic. It operates on split level sinusoidal clocks V_{CLK} and V_{CLKBAR}. Split phase clocks owing to their low voltage span of $V_{DD}/2$ over complementary sinusoidal clocks having a span of V_{DD} ensure slow charging/discharging for lower power dissipation. Figure 1(a) shows DFAL Inverter structure and (b) shows the output waveform. The structure is similar to that of a CMOS Inverter, except for two split level sinusoidal clocks V_{CLK} and V_{CLKBAR} and transistor employed diode through MT, which is used for recycling the charges. DFAL dissipates power mainly in form of adiabatic losses due to the ON resistance of the transistor MT unlike the non-adiabatic loss in diode based logic families.

$$V_{CLK} = 3/4V_{DD} + V_{DD}/4\sin(wt + \Theta) \tag{7}$$

$$V_{CLKBAR} = 1/4V_{DD} - V_{DD}/4\sin(wt + \Theta) \tag{8}$$

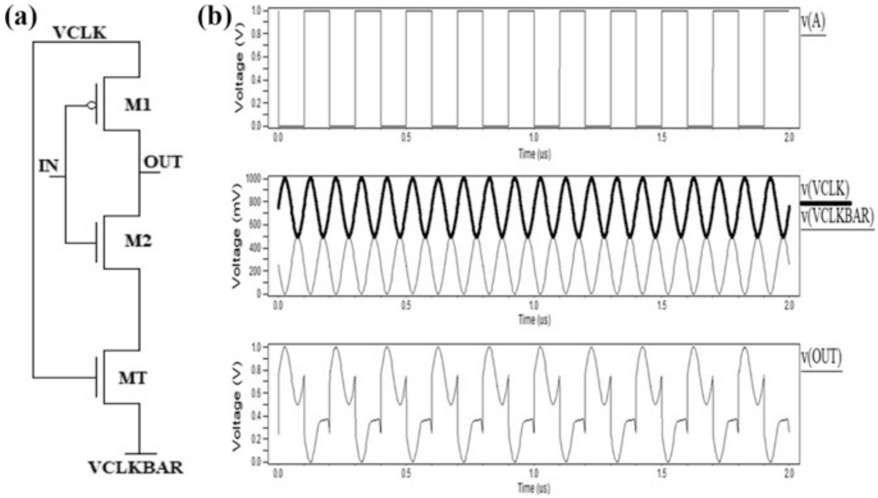


Fig. 1 DFAL inverter and waveform

3 Single Clock Diode Based Adiabatic Logic

3.1 Single Clock Diode Based Adiabatic Logic

The proposed logic is illustrated through an inverter in Fig. 2. It may be observed that the charging path is similar to DFAL in the sense that a triangular split level is employed for charging process. However, it uses a diode (MD) in its discharging path, which reduces the discharging rate (the current flow), thus the adiabatic losses given by Eq. (6). The power clock is described by (9) and (10).

$$V_{CLK} = 0.5 * V_{DD} + (V_{DD} * t)/T, \; 0 < t < T/2 \tag{9}$$

$$= 1.5 * V_{DD} - (V_{DD} * t)/T, \; T/2 \le t \le T \tag{10}$$

The clock operation can be divided into two phases—Evaluation and Hold Phase. In evaluation phase, V_{CLK} ramps up while it ramps down in a hold phase. Following is the operation:

- In evaluation phase, the load capacitance is charged according to V_{CLK} when output node is LOW and transistor M1 is turned ON. When output node is HIGH and transistor M2 is turned ON, load capacitance discharges towards the ground via the diode and PDN.
- In the hold phase, when the output node is HIGH and M1 is ON, the output node follows the clock, while when output node is LOW and PDN is ON, the output node stays at logic LOW in floating condition. When the charge on the node falls below the threshold voltage of diode, it goes into floating state. Although it is in floating state, whenever the output node voltage tries to reach above the threshold voltage of diode, due to any circuit transient, the diode will switch ON and pull back the node into floating condition. It can be thought of as negative feedback mechanism, wherein diode will ensure node voltage always remains in floating condition towards logic LOW.

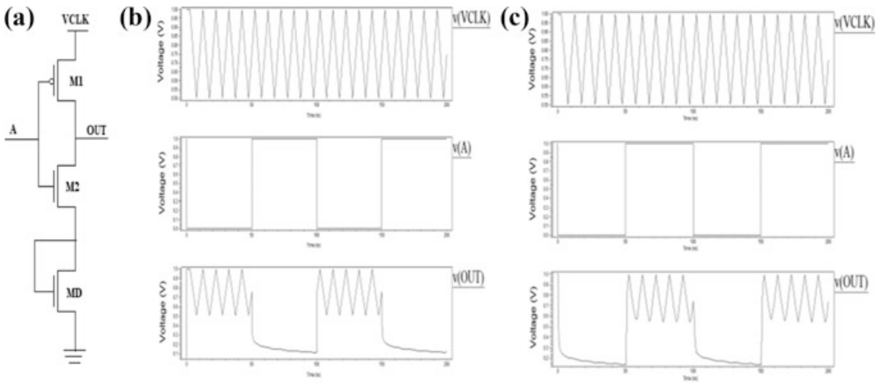


Fig. 2 SC-DBAL **a** Inverter **b** Inverter Waveform **c** Buffer Waveform

The proposed inverter is simulated using TANNER EDA Tools in 45 nm High Performance Technology and corresponding input/output waveforms are depicted in Fig. 2. It may be noted that when output is in logic HIGH state, it follows the power clock. When the output node is logic LOW, it will be in floating condition, but the presence of the diode ensures that the output node never goes high above the diode threshold. Therefore, the simulation results adhere to the functionality.

With the increase in scaling, the supply voltage has gone down, leading to reduction in the voltage difference between any two nodes in the circuit. Non-adiabatic losses, arising due to the voltage difference between the nodes, too have gone down. But adiabatic losses depend on the resistance of the path, which depends on $V_{OV} = (V_{GS} - V_{TH})$. With scaling, this overdrive voltage does not remain constant; rather it decreases because V_{TH} is not falling as fast as $V_{GS,}$ leading to an increase in adiabatic losses.

The single clock usage in the proposal as opposed to dual clocks in DFAL is advantageous, as the clock generation circuitry in the DFAL circuit not only needs to generate two separate clocks V_{CLK} and V_{CLKBAR}, but also has to maintain the phase relation between the two clocks. Failure to maintain correct phase relationship will lead to more power dissipation than expected. For the proposed logic, a single clock needs to be generated, eliminating the phase matching requirement. Hence, the proposed circuit, reduces the discharging rate and the reduction of adiabatic losses is more power efficient than DFAL, which saves non-adiabatic losses. The proposed logic is advantageous for lower technology node and operates on a single clock.

4 1 Bit Alu Design Using SC-DBAL

ALU belongs to the class of combinational circuits for performing logic and arithmetic operations. Figure 3 shows the design of 1 bit ALU for performing 8 basic operations on the two input operands, which are listed in Table 1.

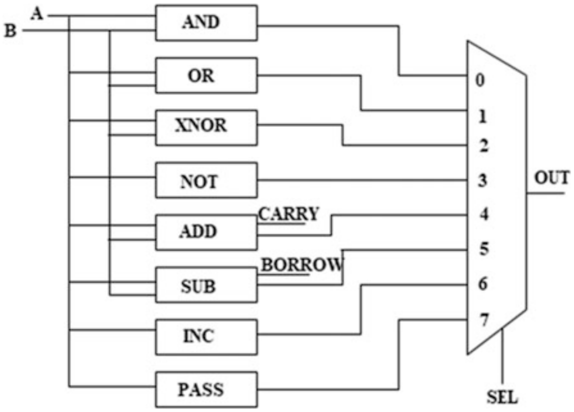


Fig. 3 1 bit ALU

Table 1 ALU truth table

Sel	Functionality	Response
000	AND	Fig. 4(c)
001	OR	Fig. 5(c)
010	XNOR	Fig. 6(b)
011	One’s complement	Fig. 2(b)
100	ADD	Fig. 7(c)
101	Subtract	Fig. 7(b)
110	Increment	Fig. 7(a)
111	Pass	Fig. 2(c)

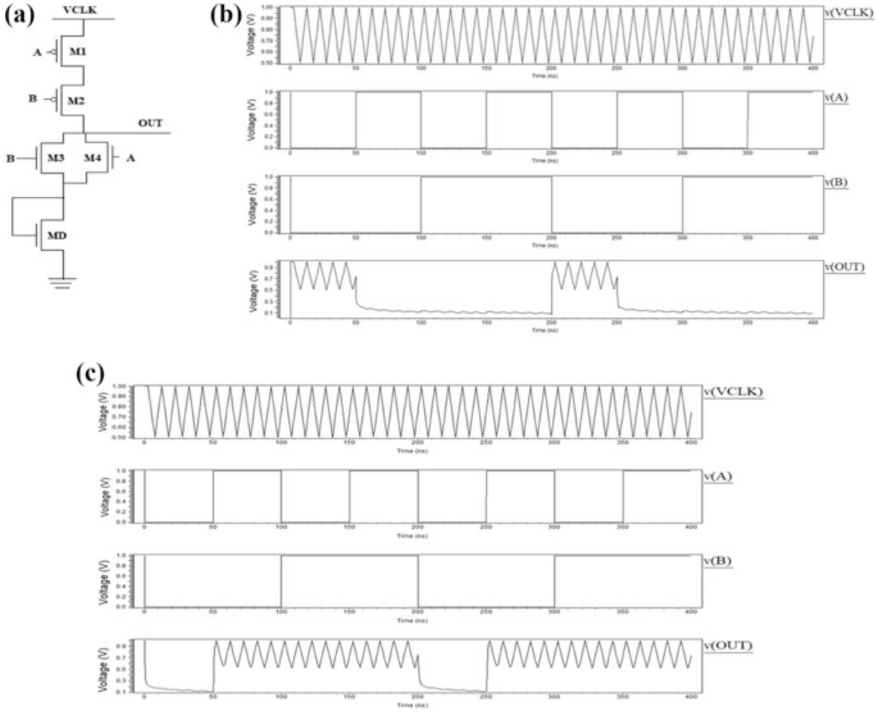


Fig. 4 SC-DBAL **a** NOR Gate and **b** NOR Waveform **c** OR Waveform

The Arithmetic Unit comprises of adders, incrementors, and subtractors, while Logical Unit of an ALU can be realized using basic gates like AND, OR, XOR and Inverter.

Figure 4(a) shows the design of two-input SC-DBAL NOR logic. Figure 4(b) and (c) show the simulation waveforms pertaining to a two-input SC-DBAL NOR and OR logic respectively and adhere to the respective functionality.

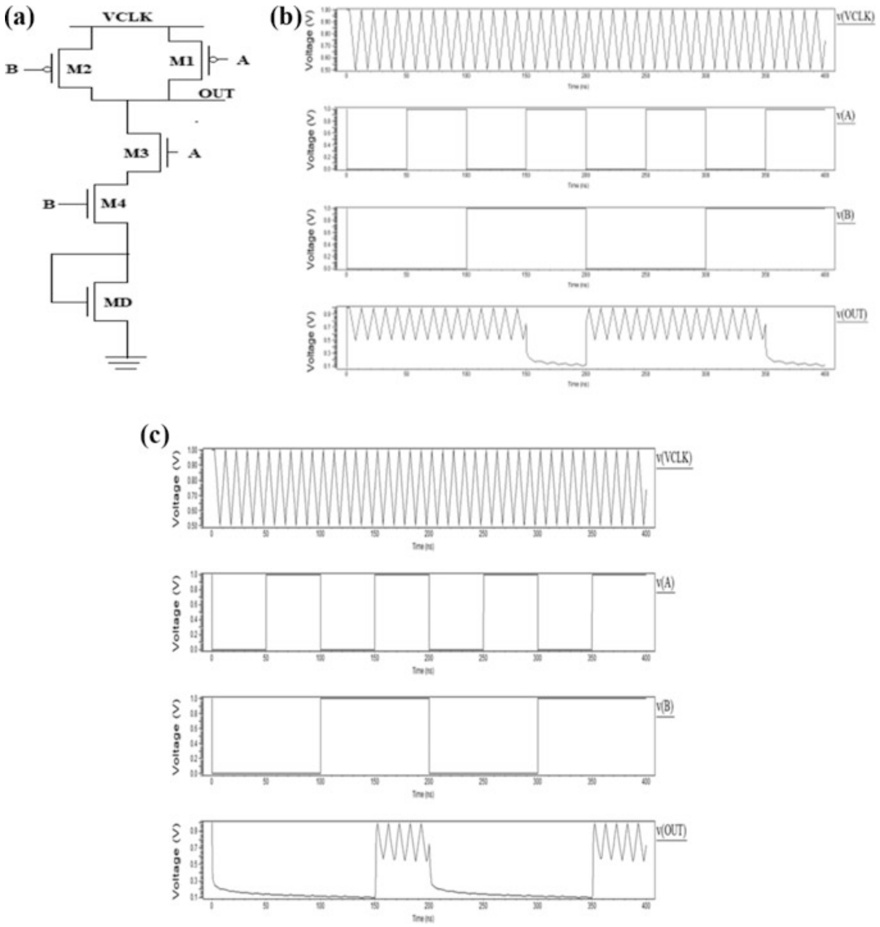


Fig. 5 SC-DBAL **a** NAND Gate **b** NAND Waveform **c** AND Waveform

Figure 5 shows the design of two-input SC-DBAL NAND logic. Figure 5(b) and (c) show the simulation waveforms pertaining to a two-input SC-DBAL NAND and AND logic respectively and adhere to the respective functionality.

Fig. 7(a)–(c) show the waveforms pertaining to SC-DBAL incrementor, subtractor and adder respectively designed using SC-DBAL circuit and adhere to the respective functionality.

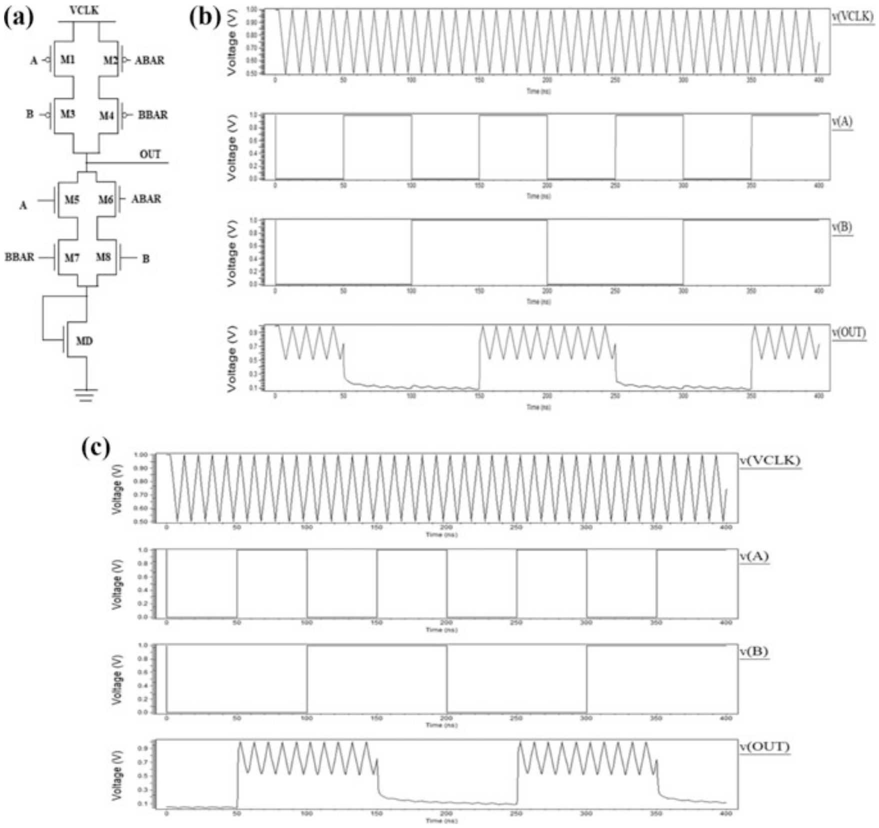


Fig. 6 SC-DBAL **a** XNOR Gate **b** XNOR Waveform **c** XOR Waveform

5 Simulative Investigation

5.1 Simulation Environment

The Adiabatic Logic Circuits use time varying voltages called power clocks—triangular or sinusoidal shaped as supply voltages and inputs. Hence, it becomes imperative that realistic supplies and inputs be used for simulating the Design Under Test (DUT) in order to correctly observe the circuit's behavior for the power estimation from the simulation tool, be as much close to real scenario as possible. To perform realistic analysis, a standardized simulation setup [4] has been established as shown in Fig. 8. It uses two stages of buffer before the DUT for input wave shaping and two stages after the DUT for imitating practical loads in a digital circuit. The power measurement of a DUT consisting of M - inputs and N - outputs is shown. It shows that the power measurements for any N input, M-output

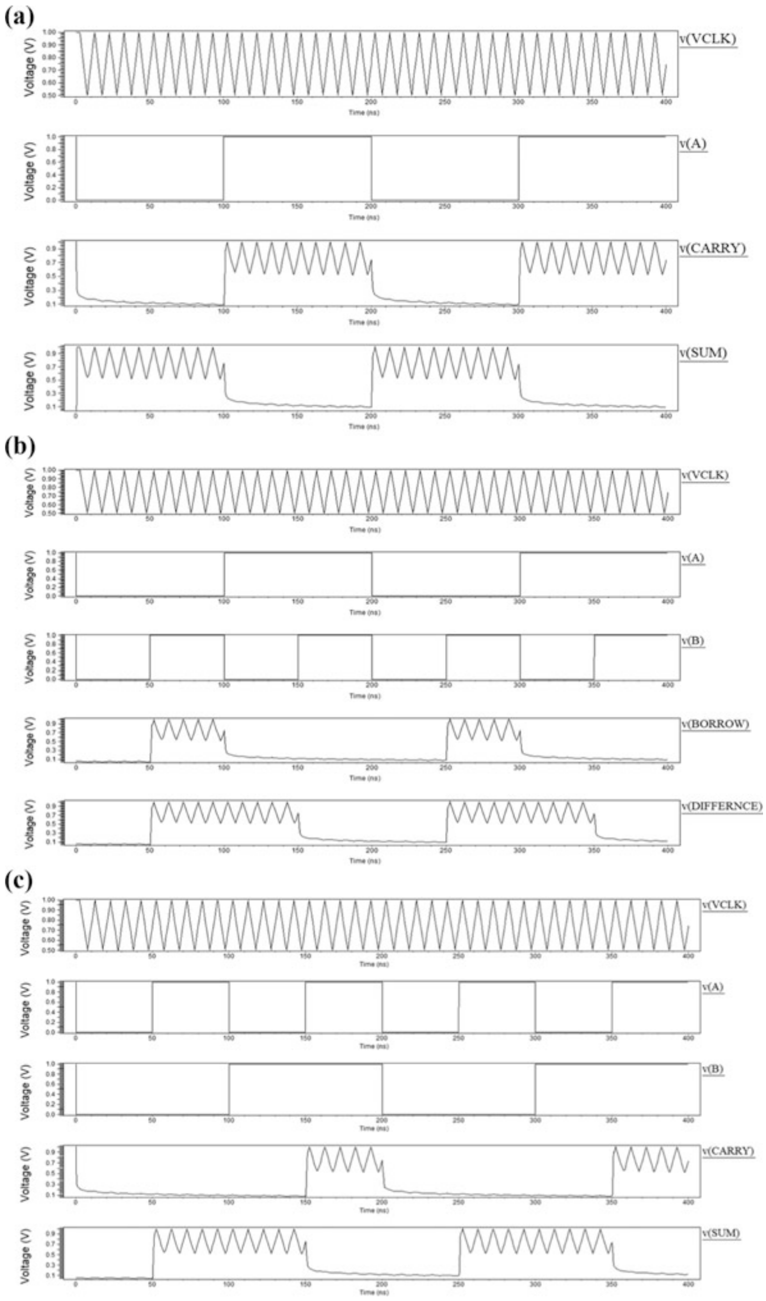


Fig. 7 **a** Increment waveform **b** Subtraction waveform **c** Addition waveform

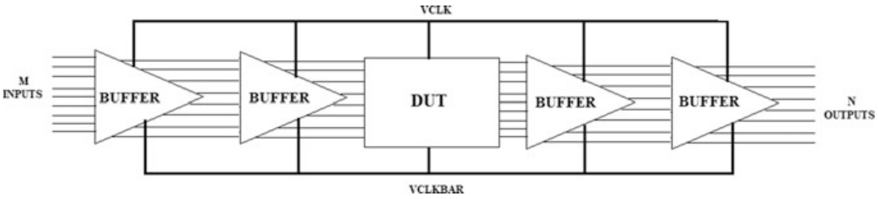


Fig. 8 Simulation test bench

Table 2 Simulation settings

Technology	PTM 45 nm HP
EDA tool	Tanner EDA tools
MOS dimension	W_{PMOS} = 300 nm, W_{NMOS} = 100 nm, W_{MD} = 450 nm
Maximum supply voltage	1 V
Power clock frequency	Two times of Input

adiabatic circuit requires ideal signals, be applied to the first stage, whereas the DUT is fed by outputs from the second stage. Table 2 shows the simulation settings used for performance comparison of the circuits presented in this paper.

For DFAL, V_{CLKBAR} is split level sinusoidal clocks, whereas for Proposed and CMOS it is ground in the Simulation Testbench.

5.2 *Power Measurement*

5.2.1 Variation with Frequency

Figures 9 and 10 show the variation of power dissipated with frequency by the Proposed, DFAL and CMOS for Single Stage and an 8 Stage Inverter Stage. Figures show that the Proposed Logic dissipates the least power compared to DFAL and CMOS for different values of frequencies.

Figures 11 and 12 show the variation of power dissipated for the Proposed, DFAL and CMOS by simulating NAND and NOR Logic respectively. Figures show that the proposed logic dissipates least power compared to the DFAL and CMOS for different values of frequencies.

Figure 13 shows the variation of power dissipated for the Proposed, DFAL and CMOS by simulating XOR gate. Figure shows that the proposed logic dissipates the least power compared to DFAL and CMOS for different values of frequencies.

All the above figures show that the proposed logic shows more savings over DFAL and CMOS for higher values of frequencies. This is in quite agreement with the fact that as frequency of operation is increased, the adiabatic losses dominate

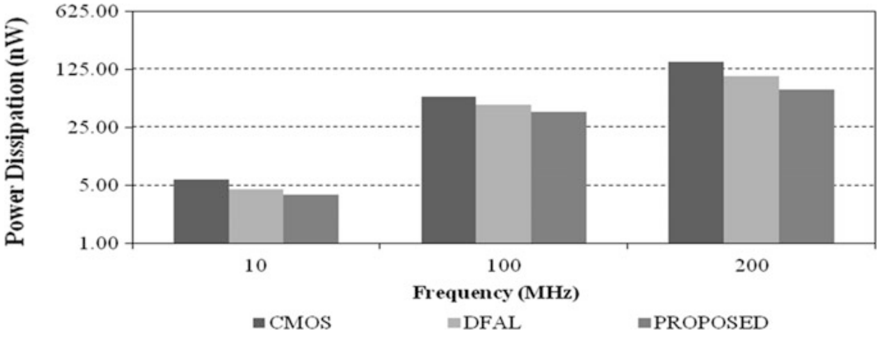


Fig. 9 Variation of power dissipation for an Inverter with Frequency

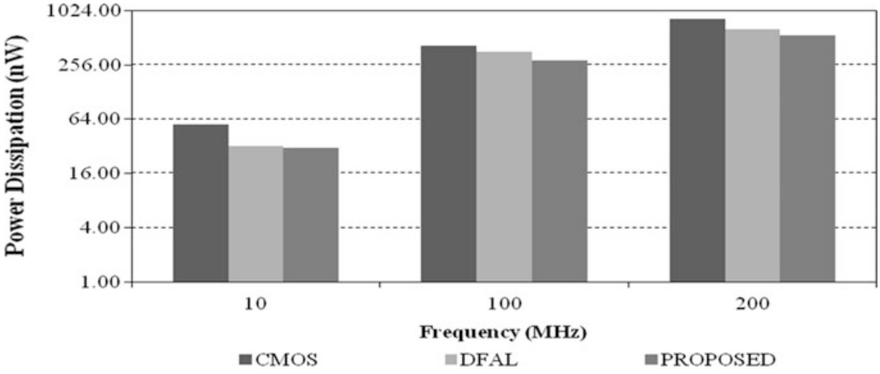


Fig. 10 Variation of power dissipation for an 8 stage Inverter with Frequency

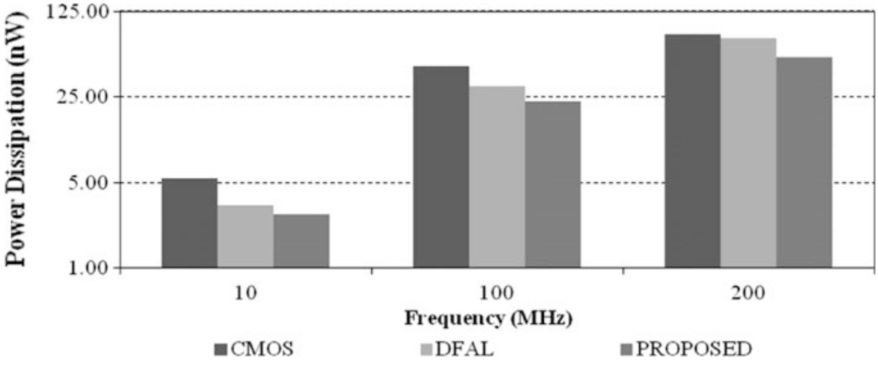


Fig. 11 Variation of Power Dissipation for a NAND Gate with Frequency

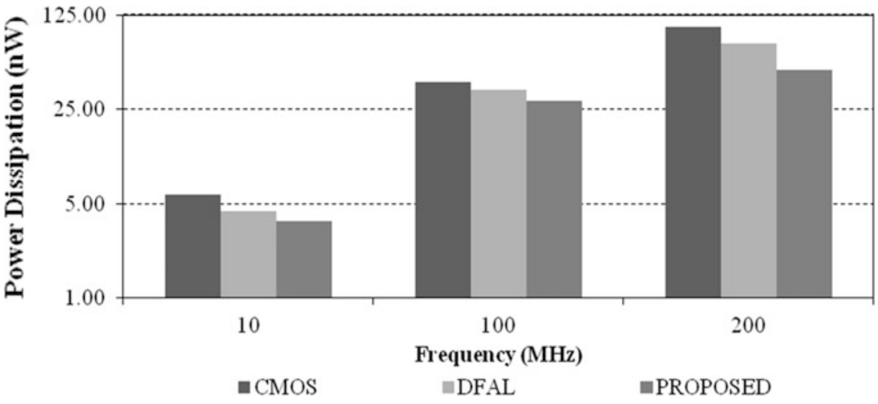


Fig. 12 Variation of power dissipation for a NOR Gate with Frequency

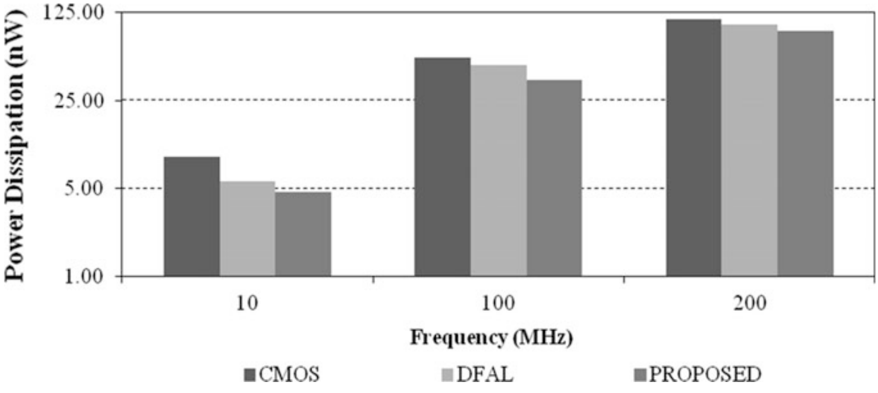


Fig. 13 Variation of power dissipation for an XOR Gate with Frequency

over the non- adiabatic losses. DFAL reduces the non-adiabatic whereas the proposed logic reduces adiabatic losses, so higher power savings are observed at high frequencies.

5.2.2 Variation with Capacitance

Figure 14 shows the variation of power dissipation for different values of capacitances. It shows that the proposed logic dissipates least power in comparison to DFAL and CMOS logic for all values of capacitances.

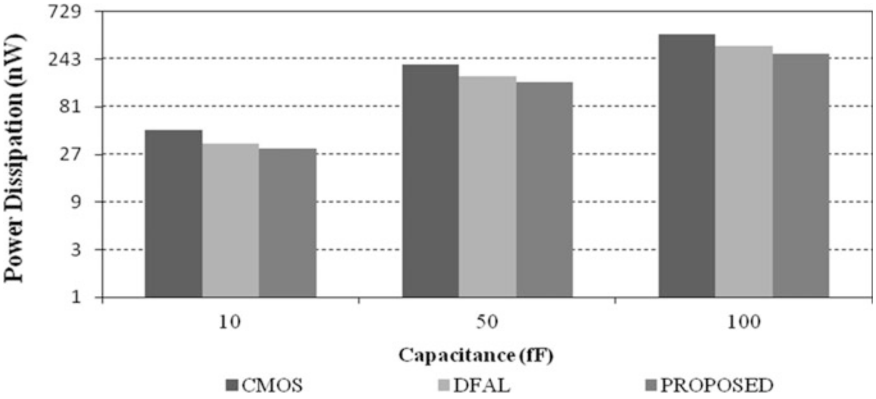


Fig. 14 Variation of power dissipation for an Inverter with Capacitance

6 Conclusion

In this paper, we proposed a new adiabatic logic family SC-DBAL (Single Clock Diode Based Adiabatic Logic Family). The proposed logic uses single clock in its charging path and diode in its discharging path for reducing the discharge rate for reducing the adiabatic losses. Simulations have shown that the proposed logic dissipates least power among the existing state of the art adiabatic families in terms of power. Basic Gates designed for 1 bit ALU also show the same trend of lower power dissipation over DFAL and CMOS. In terms of circuit complexity, it uses a single clock, lessening the burden on the clock generation circuitry.

References

1. Jain S, Pandey N, Gupta K (2018) Complete charge recovery diode free adiabatic logic. In: 2018 5th International conference on signal processing and integrated networks (SPIN). IEEE, pp 656–660
2. Hodges DA, Jackson HG Saleh RA (2003) Analysis and design of digital integrated circuits: in deep submicron technology, 3rd ed. McGraw-Hill higher education, Boston (Mass.)
3. Grover V, Gosain V, Pandey N, Gupta K (2018) Arithmetic logic unit using diode free adiabatic logic and selection unit for adiabatic logic family. In: 2018 5th international conference on signal processing and integrated networks (SPIN). IEEE, pp 777–781
4. Teichmann P (2012) Adiabatic logic in future trend and system level perspective. 1st edn. Springer Series in Advance Microelectronics
5. Dickinson AG, Denker JS (1995) Adiabatic dynamic logic. IEEE J Solid-State Circuits 30:311–315
6. Bindal V (2016) Adiabatic logic circuit design. IJISET—Int J Innov Sci Eng Technol 3:688–694
7. Reddy NSS, Satyam M, Kishore KL (2008) Glitch free and cascadable adiabatic logic for low power applications. Asian J Sci Res 1(4):464–469

8. Kim S, Papaefthymiou MC (1999) Single-phase source coupled adiabatic logic. International symposium on low power electronics and design
9. Upadhyay S, Nagaria RK, Mishra RA (2013) Low-power adiabatic computing with improved quasi static energy recovery logic. VLSI Design, 9 p
10. Vetuli A, Pascoli S, Reyneri LM (1996) Positive feedback in adiabatic logic. Electron Lett 32:1867–1869
11. Upadhyay S, Mishra RA, Nagaria RK, Singh SP (2012) DFAL: diode free adiabatic logic circuits. ISRN Electronics, vol 2013, 12 p
12. Gong C-SA, Shiue M-T, Hong C-T, Yao KW (2008) Analysis and design of an efficient irreversible energy recovery logic in 0.18-µm CMOS. IEEE Trans Circuits Systms-I: Regul Pap 55(9):2595–2607
13. Ye Y, Roy K (2001) QSERL: quasi-static energy recovery logic. IEEE J Solid-State Circuits 36(2):239–248
14. Upadhyay S, Mishra RA, Nagaria RK, Singh SP, Shukla A (2013) Triangular power supply based adiabatic logic family. World Applied Sciences Journal 24(4):444–450

Optimization of Regressing Analysis Technique to Estimate the Number of Tourists in Popular Provinces in Tourism of Thailand

Prapai Sridama

Abstract The objective of this research is to estimate the number of tourists in the provinces of Thailand with the international tourists. This research develops the Optimization of Regressing Analysis (ORA) model, which is used to estimate number of international tourists in each province of Thailand. Many mathematics equations are used to find the results of the ORA model that are linear regression for estimated number of international tourists, the first derivative for to find the trend of increased values, and the second derivative for to find the increase in international tourist arrivals. Results of this experiment show that the province with the highest number of tourists is Bangkok and the model can estimate the number of tourists are closest, including Bangkok, Chiang Mai and Phuket that three provinces with the highest number of tourists too. In addition, the average number of visitors from all provinces is 2.65%.

Keywords Linear regression · First derivative · Second derivative

1 Introduction

Over the last 4–5 years, apart from exports, tourism is a major economic driver of the country. In some years, tourism growth is half of the growth rate of the economy ever. Many provinces in Thailand can earn revenue from tourism such as Bangkok, Phuket, Chonburi Chiang Mai etc. In 2017, the top three provinces with the highest revenue from tourism are Bangkok with estimated income 17128410.81 USD, Phuket 10982011.54 USD and Chonburi with income of 5214.49 USD. In addition, consumption by foreigners accounts for about 12% of total domestic

P. Sridama (✉)
Department of Computer Science, Faculty of Science and Technology,
Bansomdejchaopraya Rajabhat University, Bangkok 10600, Thailand
e-mail: prapaikmutnb@gmail.com

Z. Zakaria and R. Ahmad (eds.), *Advances in Electronics Engineering*,
Lecture Notes in Electrical Engineering 619,
https://doi.org/10.1007/978-981-15-1289-6_24

consumption, which is considered to be a fast growing component. And become an important machinery of the Thai economy.

Current purchasing power in the country is slow. Household debt rises. Labor market slowdown and the price of agricultural products down. These causes are the impact on the growth of the Thai economy. Then the government is trying to stimulate the Thai economy with tourism policy. Tourism measures have been highly successful in terms of revenue. In 2016, Thailand earned 7657247.03 USD in tourism revenue. TMB analytics said that the Thai economy in 2018 is expected to grow 4.2%, up from the previous estimate of 3.8%, in line with the stronger global economic growth. The growth rate of the Thai economy was 4.2% from the tourism sector of 2%, reflecting that tourism is still a service sector that plays a key role in generating economic growth for the country.

The number of tourists is increasing rapidly, so the main tourist attractions cannot accommodate tourists such as accommodation facilities. And environmental problems as well (Thorn 2018). However, there are the number of issues that Thailand should consider in order to make sustainable tourism and income distribution to Thai people across the country. The reason is that foreign tourists still have some tourism. Thailand also has secondary tourism support. In addition, the government has taken measures to deal with the wave of tourists. To reduce the impact of tourist with a good host. To maintain the relationship of the country by providing visitors with information and rules in the form of a brochure or CD. Moreover, Tourism Authority of Thailand has proposed measures to build confidence and stimulate the Thai tourism market. Development of security system for tourists. Visa exemption for tourists in high growth markets and the visa to enter many times for surfers [1].

For the above reasons, the researcher sees the importance of the tourism industry and the problem of adequate tourism. Researcher is interested in conducting research on to forecast of the number of foreign tourists in the province that can develop into the provinces that provide tourism services. In addition, the purposes of this research are to forecast the number of tourists in the provinces of Thailand and makes the decision to plan accommodation and facilities enough to continue to tourists.

In this research, the researcher selected the regression analysis technique to be used to find the relationship of the variables [2]. In addition, the researcher used a simple linear regression analysis technique to analyze the regression of one independent variable and one variable [3]. Moreover, the first derivative and the second derivative are used to increase the efficiency of the simple linear regression analysis technique too.

In the next sequence is the description of the theory and related research. Method of conducting research, results of experiment and conclusion of experiment. This will be described as follows.

2 Literature Review

This past presents theories and related researches that theories can show in first past and related researches are shown in next title.

2.1 Linear Regression

Linear regression explained how the mean values of a response variable (the symbol is Y) vary as a linear function of a set of explanatory variables. For example, there is only a single explanatory variable, denoted by X; when restricted to a single explanatory variable the model is referred to as simple linear regression [4].

Assume that there are observations on individuals $(i = 1, \ldots, n)$ in the study, the simple linear regression model is

$$Y_i = \beta_0 + \beta_1 X_i + e_i \tag{1}$$

where the β's are the regression coefficients and e_i is a random value that is assumed to be independently normally distributed with zero mean and variance σ^2. The e_i term in the simulation represents natural variation of Y_i among individuals around the mean or expected value of the response in the population,

$$E(Y_i) = \beta_0 + \beta_1 X_i \tag{2}$$

Regression coefficients, β_0 and β_1, express the linear dependence of the mean response on the explanatory variable. The intercept β_0 has interpretation as the mean value of the response variable when the explanatory variable X is equal to zero. β_1 is the parameter of most interest and has interpretation as the change in the mean of Y for a single unit increase in X. In the special case, the explanatory variable X is dichotomous, taking values of 0 and 1, the regression slope β_1 has a simple interpretation as the difference in the mean of Y when X = 1 versus X = 0.

2.2 The Second Derivative

The interpretation of the second derivative can be used to test whether a stationary point. The stationary point of a differentiable function of one variable is a point on the graph of the function where the function's derivative is zero [5, 6]. Normally, this is a point where the function "stops" increasing or decreasing. Stationary points are easy to visualize on the graph of a function of one variable: they correspond to the points on the graph where the tangent is horizontal. For a function (i.e. a point where $f'(x) = 0$ is a local maximum or a local minimum).

If $f''(x) < 0$ then f has a local maximum at x point. If $f''(x) > 0$ then has a local minimum at point. If $f''(x) = 0$, a possible inflection point.

2.3 Related Researches

Developed goods characteristics theory to learn product decision making. This assume that utility is pursued from the characteristics or objects of a product. The theory fits the tourism context, since destinations consist of a range of intangible and tangible attributes, including social, cultural and environmental features [7]. However, tourists do not derive utility by possessing or using destinations as a whole; they success utility by consuming specific destination components such as transport, accommodation and attractions [8]. Lancaster's goods characteristics theory has been used widely in tourism researches to identify the factor of destination choice, and over the years this theory has been adopted and refined [9–13]. However, the increased accuracy of prediction about tourist's destination choice. A large number of studies have estimated tourist's preferences based on this approach. These researches can be categorized into two types that are revealed-preference estimation [14] and the stated-preference estimation [15–18].

3 Methodology Materials

The optimization of regressing analysis technique to estimate the number of tourists in popular provinces in tourism of Thailand is developed by a model (Optimization of Regressing Analysis: ORA). In addition, the ORA model is explained in this section that procedures of the ORA model can be shown below (see Fig. 1).

There are equations (linear regression, first derivative, and second derivative) f statistics mathematical for using to forecast the number of tourists.

Figure 1 shows procedures of the ORA model chart. The first step, the ORA model loads data of the number of tourists in each province of Thailand into this model. Then the first derivative equation is used to find the increasing trend or decreasing trend. If provinces are decreasing trends then these provinces are removed to consider from the ORA model. The second step, to find the rate of increasing trend from provinces, which are increasing trends by the second derivative equation. The last step, the ORA model considers the results from the second derivative equation. If this equation gives high growth rate then the ORA model estimates the number of the tourists by linear regression.

In addition, the efficiency of the ORA model can be shown by to compare between the number of tourists by the ORA model and the real number of tourists.

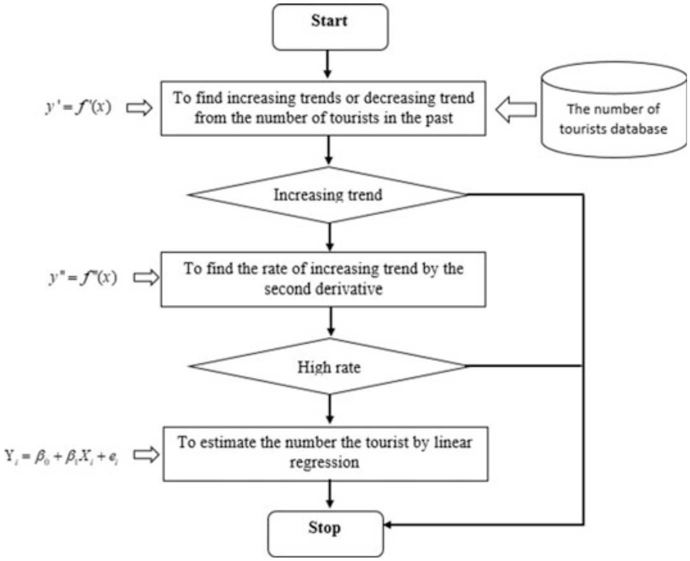

Start
$y' = f'(x)$
To find increasing trends or decreasing trend from the number of tourists in the past
The number of tourists database
Increasing trend
$y'' = f''(x)$
To find the rate of increasing trend by the second derivative
High rate
$Y_i = \beta_0 + \beta_1 X_i + e_i$
To estimate the number the tourist by linear regression
Stop

Fig. 1 Procedures of the ORA model

4 Results and Findings

This research focuses on provinces, which are with high tourist arrivals. The provinces and the number of tourists in each province are shown in Table 1.

Table 1 presents the number of international tourists between 2009 and 2015 in each province, which is tourism authority. There are five provinces, including Bangkok, Chonburi, Chengmai, Phuket, and Krabi that international tourists more than one million tourists in any years. In addition, the first derivative is used to find the trends of the number of international tourists and the second derivative is used to find the rate of increase in international tourists. The results of these equations found that there are seven provinces with higher tourist arrivals namely Bangkok, Chengmai, Phuket, Songkla, Suratthanee, Krabi, and Phangnga. Moreover, this research estimates the number of international tourists in 2016 by linear regression, that results are shown in Table 2.

Table 2 presents the number of international tourists in 2016 and the estimation of the number of international tourist in 2016 by the ORA model. The results of experiment found that the ORA model gives the estimation values is higher than the real numbers, including Bangkok, Chengmai, Phuket, Songkla, and Suratthanee. In addition, estimated results do not exceed 10% of actual values. However, there are two provinces (Krabi and Phangnga) that the ORA model gives the estimation values is lower than actual values. The reason is that the estimated value is lower than the actual value is the amount of surfers have increased in some periods and in some periods the rate has decreased.

Table 1 The number of international tourists between 2009 and 2015

Provinces	International tourists (Million)						
	2009	2010	2011	2012	2013	2014	2015
Ayuthaya	0.88	1.38	0.92	1.58	1.66	1.80	1.81
Bangkok	10	11	15	17	19	18	21
Chonburi	2.77	5.51	6.29	6.75	7.22	7.58	6.96
Chengrai	0.25	0.38	0.48	0.52	0.53	0.52	0.06
Chengmai	1.24	1.70	2.04	2.19	2.34	2.60	2.84
Prachuap Khiri Khan	0.53	0.48	0.68	0.87	0.92	0.93	0.97
Phuket	2.49	4.51	6.62	6.19	8.40	8.46	9.49
Rayong	0.68	0.24	0.25	0.47	0.47	0.47	0.47
Songkla	0.77	0.87	1.16	1.13	2.21	2.29	2.49
Sukhothai	0.20	0.21	0.19	0.27	0.31	0.32	0.34
Suratthanee	1.08	0.98	1.17	1.79	2.71	2.94	3.17
Nongkai	0.19	0.26	0.22	0.48	0.52	0.53	0.55
Srakaew	0.15	0.17	0.15	0.21	0.22	0.22	0.24
Krabi	1.13	1.15	1.34	1.44	2.00	2.78	3.49
Phangnga	0.42	0.50	0.62	0.83	1.32	1.66	3.00

Table 2 The estimation of the number of international tourists in 2016

Provinces	The number of international tourists in 2016 (million)	The estimation of the number of international tourists in 2016 by ORA model (million)
Bangkok	21	22.92
Chengmai	2.92	3.12
Phuket	9.50	10.98
Songkla	2.50	2.85
Suratthanee	3.25	3.65
Krabi	3.58	3.48
Phangnga	3.23	2.73

5 Conclusion

The purpose of this research is to estimate the number of tourists in the provinces of Thailand with the international tourists. The Optimization of Regressing Analysis (ORA) model is developed in this research for using estimated tool. The ORA model uses linear regression for estimated number of international tourists, the first derivative for to find the trend of increased values, and the second derivative for to find the increase in international tourist arrivals. Results of this experiment show that the province with the highest number of tourists is Bangkok and the model can estimate the number of tourists are closest, including Bangkok, Chiang Mai and Phuket that three provinces with the highest number of tourists too.

Acknowledgements Author thanks the Bansomdejchaopraya Rajabhat University for offering funded research and the staff of the research and development institute, which has facilitated the documents.

References

1. Thammasat Institute of area studies: Thailand-central Tourism Monsoon Competition, Tourism Authority of Thailand (2015)
2. Flavia F, Anisor N (2014) Analysis of the economic performance of an organization using multiple regression. In: International conference of scientific paper AFASES (2014)
3. Yuejin Z, Yebin C, Tiejun T (2014) A least squares method for variance estimation in heteoscedastic nonparametric regression. J Appl Math 1–14
4. Fitzmaurice GM (2016) Regression. Diagn Histopathol 22(7):271–278
5. Chiang AC (1984) Fundamental methods of mathematical economics, 3rd edn. McGraw-Hill, New York
6. David S, Julia S, Derek W (2011) 12 B stationary points and turning points. Cambridge 2 unit mathematics year 11. Cambridge University Press, Cambridge
7. Lancaster KJ (1966) A new approach to consumer theory. J Polit Econ 174:132–157
8. Tse TSM (2014) A review of Chinese outbound tourism research and the way forward. J China Tourism Res 11(1):1–18
9. Basala SL, Klenosky DB (2001) Travel-style preferences for visiting a novel destination: a conjoint investigation across the novelty-familiarity continuum. J Travel Res 40:172–182
10. Morley CI (1994) Experimental destination choice analysis. Ann Tourism Res 21:780–791
11. Papatheodorou A (2001) Why people travel to different places. Ann Tourism Res 28:164–179
12. Rugg D (1973) The choice of journey destination: a theoretical and empirical analysis. Rev Econ Stat 55:64–72
13. Seddighi HR, Theocharous AI (2002) A model of tourism destination choice: a theoretical and empirical analysis. Tour Manag 23:475–487
14. Agrusa J, Kim SS, Wang KC (2001) Mainland Chinese tourists to Hawaii: their characteristics and perferences. J Travel Tourism Mark 28:261–278
15. Ciná VZ (2012) Tourism marketing: a game theory tool for application in arts festivals. Tourism Econ 18:43–57
16. Hsu TK, Tsai YF, Wu HH (2009) The preference analysis for tourist choice of destination: a case study of Taiwan. Tour Manag 30:288–297
17. Suh YK, Gartner WC (2004) Preferences and trip expenditures: a conjoint analysis of visitors to Seoul, Korea. Tourism Manage 25:127–137
18. Tsaur SH, Wu DH (2005) The use of stated preference model in travel itinerary choice behavior. J Travel Tourism Mark 18:37–48

The Impact of Pre-processing and Feature Selection on Text Classification

Nur Syafiqah Mohd Nafis and Suryanti Awang

Abstract Nowadays text classification is dealing with unstructured and high-dimensionality text document. These textual data can be easily retrieved from social media platforms. However, this textual data is hard to manage and process for classification purposes. Pre-processing activities and feature selection are two methods to process the text documents. Therefore, this paper is presented to evaluate the effect of pre-processing and feature selection on the text classification performance. A tweet dataset is utilized and pre-processed using several combinations of pre-processing activities (tokenization, removing stop-words and stemming). Later, two feature selection techniques (Bag-of-Words and Term Frequency-Inverse Document Frequency) are applied on the pre-processed text. Finally, Support Vector Machine classifier is used to test the classification performances. The experimental results reveal that the combination of pre-processing technique and TF-IDF approach achieved greater classification performances compared to BoW approach. Better classification performances hit when the number of features is decreased. However, it is depending on the number of features obtained from the pre-processing activities and feature selection technique chosen.

Keywords Unstructured · High-dimensional · Pre-processing · Text classification · Feature selection

N. S. M. Nafis (✉) · S. Awang
Soft Computing & Intelligent System Research Group (SPINT), Faculty of Computer System & Software Engineering (FSKKP), Universiti Malaysia Pahang, Lebuhraya Tun Razak, 26300 Kuantan, Pahang, Malaysia
e-mail: nsyafiqahmnafis@gmail.com

S. Awang
e-mail: suryanti@ump.edu.my

Z. Zakaria and R. Ahmad (eds.), *Advances in Electronics Engineering*,
Lecture Notes in Electrical Engineering 619,
https://doi.org/10.1007/978-981-15-1289-6_25

1 Introduction

Currently, there is an increasing number of text documents worldwide that can be classified as unstructured and high in dimension. Besides, the text documents may contain noises, for example, short-forms, special expression, and notations. Hence, they are hard to manage, process, and utilize. Text classification is a process to classify collections of the text documents according to their predefined categories [1]. Yet, it is also a process to dig into the hidden knowledge or patterns from the text. Document indexing, spam filtering, language detection, and plagiarism detection are the examples of the text classification applications. The pre-processing is the key activity to extract the hidden knowledge in the text document. It has been an indispensable component in text classification. The aim of pre-processing is to identify the best feature representation for text classification [2]. Besides, it helps to reduce the number of dimensions of a textual data. Consequently, the effectiveness of text classification will increase. Tokenization, notation, and stop-word removal and stemming, are the activities involve in the pre-processing phase. The text documents will be chunks to meaningful tokens in the tokenization activity. Later, all the stop-words are removed. Commonly, prepositions, articles, and pronouns are defined as a Stop-words. There are hundreds of stop-words defined in the English Language. The text document also often consists of some special formats like number and date format, which need to be eliminated. Lastly, the process of stemming will be implemented. In this activity, each one of the words will be converted into its root-word. The duplication of the root-word will be removed.

After the pre-processing phase, the next is feature selection. Bag-of-Word (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) were selected for this study. The purpose of implementing these two feature selection techniques is to examine the impact of pre-processing and feature selection towards text classification. Later, the experiment proceeds with text document classification by employing Support Vector Machine (SVM). Lastly, each set of experiment is evaluated using several performance measures such as, accuracy, precision, recall, and F-1 measure. The rest of this paper is organized as follows; Sect. 2 will discuss the related work of the previous studies. Section 3 will describe the methodology for this paper. The details on the expected results will be explained in Sect. 4. Finally, in Sect. 5, some concluding remarks will be summarized.

2 Related Works

Many previous studies on the impact of pre-processing and feature selection towards text classification is reviewed and investigated covering multiple research areas.

Pre-processing gives a significant impact on the performance of many Natural Language Processing (NLP) tasks including plagiarism detection. Ceska and Fox

[3] investigate the influence of text pre-processing to detect plagiarism. It demonstrates how the pre-processing (stop-word removal, lemmatization, Number Replacement (NMR), Synonymy Recognition (SYR), and the Word Generalization (WG)) affects the text plagiarism. It emphasized the tokenization as the most important block for the text pre-processing. It also looks into the influence of punctuation and word-order within N-grams feature model. However, the experiment with the combination of NMR, SYR, and WG, obtained the highest score of 95.92% based on F1-measure compared to 95.68% when no pre-processing processes were deployed. However, the effect of the pre-processing to the classification performance is slightly improved. Besides, maintaining punctuation had given a negative impact. It helps to reduce the number of features that have to be analyzed, yet, the execution time is maintained.

Medical Literature Analysis and Retrieval System Online (MEDLINE) is known as a bibliographic database of life sciences and biomedical information, which covers broad research areas with thousands of articles. Hence, it is difficult for researchers to retrieve a reasonable amount of relevant articles using a simple query language interface. Automatic document classification is one of the solutions to this issue. Thus, Goncalves et al. [4] conduct a study to inspect the effects of pre-processing techniques in text classification of MEDLINE documents. It accesses the impact of combining different pre-processing activities (tokenization, special character and stop-word removal, pruning, stemming and WordNet) together with several classification algorithms, such as, Support Vector Machine, Decision Tree (j48), Bayes Network, and k-nearest neighbour (k-nn) on 3000 samples from the MEDLINE database. The highest accuracy was achieved by applying Bayes Network classifier and when the number of features was 916 (term appear less than 10 times are removed) with no stemming involved. Meanwhile, the lowest accuracy hit when k-nn classifier was applied on the dataset with 1304 feature and no pruning and stemming were included. The results proved that, it significantly improves the accuracy when the pre-processing activities are implemented and number of attributes are reduced.

Next, in the information retrieval (IR) system, Singh and Saini [1] proposed an effective tokenization approach which was based on the training vector. They claimed that the traditional approaches often fail to identify good tokenization solution when users request to group the documents according to their needs. They stated that tokenization is an integral part of the IR system, which identifies the tokens and their counts. Tokenization also assists to satisfy the user's information request more precisely and reduced search sharply. The number of tokens generated and pre-processing time are set as the evaluation for comparative analysis. The result shows that the number of generated tokens helps to reduce the storage space required and more accurate results are provided to the user. Secondly, the tokenization with stop-word removal and stemming as pre-processing activities, generate more accurate and effective tokens in context of information retrieval results, but with less processing time, 150 ms. They conclude that the time consumed for the entire tokenization process gives the impact on the performance of an IR system.

Pre-processing printed documents are extremely different to the digital documents. Optical Character Recognition (OCR) is utilized for the recognition of printed scripts. Shinde and Chougule [5] stated that the accuracy of the OCR system mainly relies on the text pre-processing and segmentation algorithm being implemented. Thus, they proposed a segmentation method for printed text image. In the proposed method, the text document is segmented into lines and words. No modification for segmentation of characters in any text document is needed when using this proposed method. This method successfully identifies total number of lines, total number of words, and number of words in a specific line from the input text document.

On the recent research of sentiment analysis, Jianqiang and Xiaolin [6] experimented text pre-processing method on 5 Twitter datasets. They create a set of experiment with two types of classification tasks and six pre-processing methods (replacing negative mentions, removing URL links in the corpus, reverting words that contain repeated letters, removing numbers, removing stop words and expanding acronyms). Two feature models, which are n-grams features and prior polarity score feature, are also implemented with four classifiers (SVM, Naïve Bayes, Logistic Regression, and Random Forest). Process of removing URLs, stop-words, and numbers gives less effect on the performance of classifiers. Besides, the processes are appropriate to reduce noise from the dataset. Meanwhile, replacing negation and expanding acronyms gave a significant improvement on the classification accuracy as well as for sentiment analysis.

Meanwhile, Al-Saffar et al. [7] had proposed Malay sentiment analysis based on combined classification approaches and Senti-lexicon algorithm. The goal of this research is to improve classification performances based on the semantic orientation and machine learning approaches. Pre-processing activities involved are text document cleaning, tokenization, normalization, and removal of stop-words on total of 2478 Malay sentiment-lexicon phrases. In the last phase of the experiment, three individual classifiers and a combined classifier are used to evaluate the classification accuracy. The experimental results proved that feature selection improves the performance of Malay sentiment analysis based on the combined classification. Nevertheless, the results rely on three factors: the features, the number of features, and the classification approach.

For multi-label document categorization, Islam et al. [8] proposed an automatic classification system for Bengali documents. A twelve predefined categories of document were utilized. The pre-processing tasks and feature selection are applied prior to the document classification. The N-gram model is applied to represent sentences in a sequence of words. Tokenization, symbol removal, stemming, and stop-word removal are implemented as pre-processing activities. TF-IDF is chosen as the feature selection technique and SVM as the classifier, and the proposed technique obtained 92.58% F-1 measure score and higher precision and recall rate.

From the above literature reviews, there are a few numbers of proper and deep analysis of the effect of the combination between text pre-processing and feature selection on the text classification. In order to fill the gap of this study, it concerned with accessing the effects of the text pre-processing on text classification using two different feature selection techniques on Twitter datasets.

3 Methodology

Firstly, we will briefly explain the pre-processing activities involved in this experiment set-up.

3.1 Pre-processing Activities

Tokenization

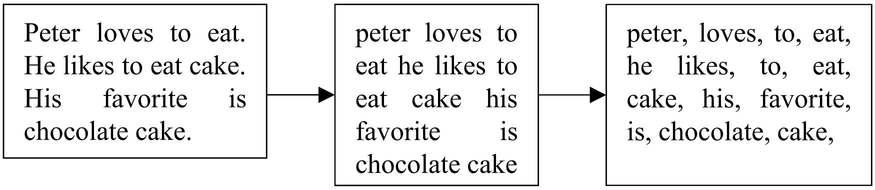
Tokenization is a process to chunk sentences into meaningful features known as token. In some cases, it is also known as a segmentation. Referring to Fig. 1, it depicts the process of tokenization. Firstly, all punctuation, notation, numbers, date format, and special characters are eliminated. Then, all words are converted into lower-case latter. Next, using whitespace as delimiter, the text document was chunked from a large paragraph or sentences into a single token. Lastly, text documents are delimit again using special character, in this case a comma.

Stop-word Removal

In stop-word removal, the aims is to clean-off tokens in the text that are generally considered as “functional word”, for example, “after”, “before”, “but”, “and” which do not carry any significant meaning for the text document. It is frequently appears in the English language. Besides, eliminating stop-word will also reduce the dimensionality. SMART stop-list is used in this research [9]. Figure 2 summarizes the algorithm to eliminate stop-words from the text document. It begins by loading tokens and stores it in array, T. Then, the process is continued by reading stop-word list one-by-one and compared with the array T using sequential search technique until matching pattern matched. Matching words will be removed from array T. These processes are continued until length of T array is met. The output is the new array tokenized word without any stop-words.

Stemming

Stemmers are used to combine tokens to optimize retrieval performance and to reduce the size of dimensionality. Stemming will identify the root of a term by removing the suffixes. Besides, singular or plural words are also stemmed. For this

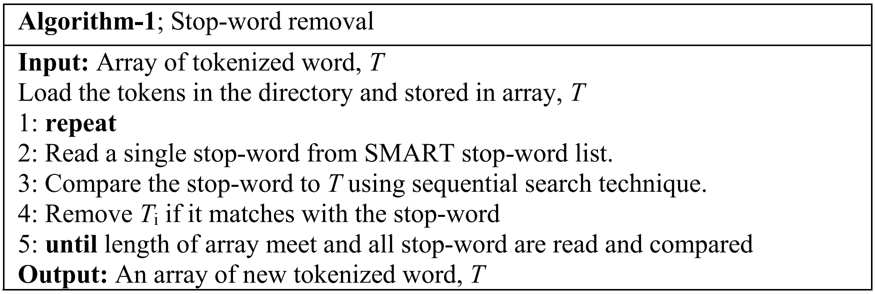


Peter loves to eat. He likes to eat cake. His favorite is chocolate cake.

peter loves to eat he likes to eat cake his favorite is chocolate cake

peter, loves, to, eat, he likes, to, eat, cake, his, favorite, is, chocolate, cake,

Fig. 1 Tokenization process



```
Algorithm-1; Stop-word removal
Input: Array of tokenized word, $T$
Load the tokens in the directory and stored in array, $T$
1: repeat
2: Read a single stop-word from SMART stop-word list.
3: Compare the stop-word to $T$ using sequential search technique.
4: Remove $T_i$ if it matches with the stop-word
5: until length of array meet and all stop-word are read and compared
Output: An array of new tokenized word, $T$
```

Fig. 2 The example of before-after removing stop-words

Table 1 Porter stemmer group of conditions

Condition No.	Condition	Description
1	m	Measure m = k or m > k, where k is an integer
2	*X	The stem ends with a given letter X
3	*v*	The stem consists of a vowel
4	*d	The stem ends in double consonant
5	*o	The stem ends with a consonant-vowel-consonant sequence, where the final consonant is not w, x or y

research, Porter's algorithm for English sentences [10] is selected since it is commonly used for many text classification purposes due to its effectiveness. Generally, Porter stemmer can be defined as follows;

$$[C](VC)m[V] \tag{1}$$

where, C and V are lists of consonants and vowels, respectively, and m is the measure of the word. Porter stemmer utilized about sixty rules, divided into 5 groups of conditions to accurately derive the stem of a word as in Table 1. Basically, rules can be written in the form of, (condition) S1 → S2 where S1 and S2 are suffixes. In a given set of rules, only the one with the longest matching suffix S1 is applied. Besides, rules are also divided into several steps to generate a stemmed word. Figure 3 summarizes Porter stemming steps based on rules explained in Table 1.

3.2 *Feature Selection*

In order to evaluate the impact of different pre-processing activities and feature selection techniques, two feature selection technique are chose. This section will

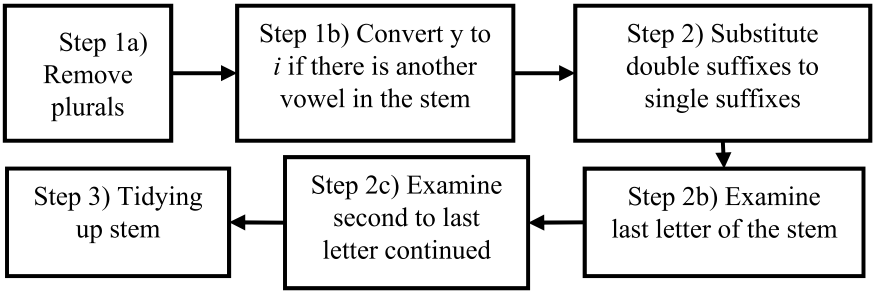

Step 1a) Remove plurals
Step 1b) Convert y to *i* if there is another vowel in the stem
Step 2) Substitute double suffixes to single suffixes
Step 2b) Examine last letter of the stem
Step 2c) Examine second to last letter continued
Step 3) Tidying up stem

Fig. 3 The Porter stemmer condition steps

explain the methodology used for Bag-of-Word first, and later, are the details on TF-IDF methodology.

Bag-of-Word (BoW)

BoW is a conventional approach, which is commonly employed in Natural Language Processing (NLP) and Information Retrieval (IR) to select features [11]. It considers a text document to be a series of independent features [12]. Hence, it neglects the ordering and structure of the text document. Given three samples of documents, (D1) Peter loves to eat, (D2) He likes to eat cakes, and (D3) His favorite is chocolate cake and fruit cake. Firstly, for every unique words generated, they are presented in document-term frequency as shown in Fig. 4.

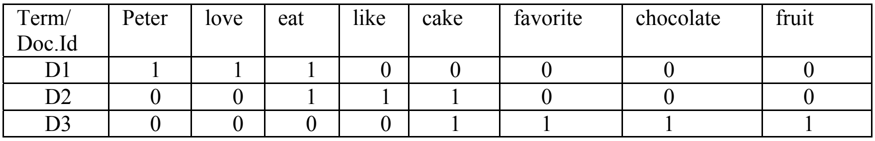
Then, BoW chooses certain number of words for classification. For example in high-dimensional data, BoW may choose top 1000 frequent words and set these 1000 features for classification. Later, a document vector based on present (1) and absent (0) is created as Fig. 5.

Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF is a filter feature selection approach and the most well-known feature weighting technique, especially in Information Retrieval system. It calculates the importance and relevance of a feature of a large document collection. TF-IDF defines that the feature relevance increases proportionally with the frequency of a word appears in a document compared to the inverse proportion of the same word in the whole collection of documents. Term frequency (TF) represents the total number or terms that appear in a document, while inverse document frequency (IDF) determines the importance of a term. Details on the formula are written as;

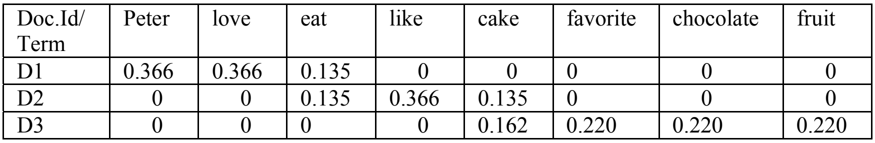
Term/ Doc.Id	Peter	love	eat	like	cake	favourite	chocolate	fruit
D1	1	1	1	0	0	0	0	0
D2	0	0	1	1	1	0	0	0
D3	0	0	0	0	2	1	1	1

Fig. 4 The document-term frequency



Term/ Doc.Id	Peter	love	eat	like	cake	favorite	chocolate	fruit
D1	1	1	1	0	0	0	0	0
D2	0	0	1	1	1	0	0	0
D3	0	0	0	0	1	1	1	1

Fig. 5 The document vector



Doc.Id/ Term	Peter	love	eat	like	cake	favorite	chocolate	fruit
D1	0.366	0.366	0.135	0	0	0	0	0
D2	0	0	0.135	0.366	0.135	0	0	0
D3	0	0	0	0	0.162	0.220	0.220	0.220

Fig. 6 The document vector based on TF-IDF

$$TF\text{-}IDF(t, d, D) = TF(t, d) * IDF(t, D) \tag{2}$$

where, let D = {d1, d2, d3, …, dn} be a collection of documents and t be a term that occurs in the collection. TF(t, d) represents the frequency of the term t in the document d. Meanwhile, IDF(t, D) is the inverse document frequency, where t represents the frequency of the term that appear in D. D is the number of the document in the collection. The example of TF-IDF calculation is briefly explained as follows: given three document samples, (1) Peter loves to eat, (2) He likes to eat cake, and (3) His favorite is chocolate and fruit cake. TF-IDF also will calculate the document-term frequency as in BoW. However, what makes TF-IDF vector different is the IDF value calculated.

By using formula 2, it will produce document vector as in Fig. 6. The feature is more representative if it has larger TF-IDF value. In this case, token Peter, love and like has the highest TF-IDF value, which indicates that these three tokens are the most prominent features all across the document collection and lower scores terms of TF-IDF can be eliminated based on some thresholds. These thresholds are varied according to the numbers of features processed (token). Figure 7 illustrates the process flow of this experimental setup. Hence, it generates six set of experiments as shown in Table 2.

4 Result and Discussion

The main focus of this research is to study the impact of the pre-processing and feature selection on the English text document. A collection of tweets is selected. It comprises of 150 tweets retrieved from an established repository. The tweets have been annotated as 0 = negative and 1 = positive. 72 samples are labelled as

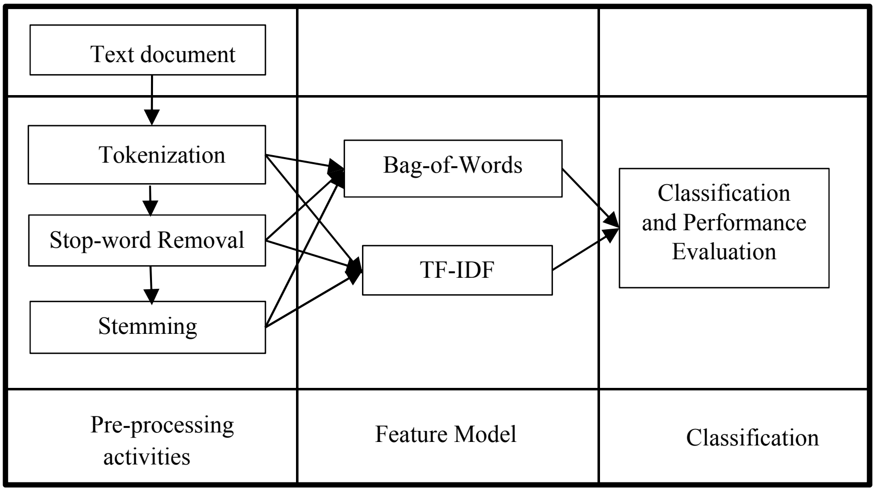


Fig. 7 The process flow

Table 2 The series of experiments

Experiment	Description
1	Text classification with tokenization only and BoW
2	Text classification with tokenization, stop-word removal, and BoW
3	Text classification with tokenization, stop-word removal, stemming and BoW
4	Text classification with tokenization only and vector TF-IDF
5	Text classification with tokenization, stop-word removal, and TF-IDF
6	Text classification with tokenization, stop-word removal, stemming and TF-IDF

negative and 78 are positive samples. The selected sample is split into 80% of training (120 sample) set and (30 samples) for testing. Experiments are demonstrated using Matlab software packages with 5 folds of cross validations during the text classification. In order to evaluate the performances of each proposed methodology, Support Vector Machine classifier is applied and four performance measures (accuracy, precision, recall, and F-measure) are accessed. Precision measures the percentage of the document that is correctly classified as positive out of the entire document, which is marked as positive. Meanwhile, recall is the percentage of the documents that are correctly classified as positive out of the entire document that is actually positive. F-measure is the combination of precision and recall to formulate one single measure. The higher the values of precision, recall, accuracy, and F-1 measure obtained during classification process, they indicates that the proposed techniques are able to improve the classification effectiveness.

Experiments 1–3 are conducted to evaluate the effects of three combinations of pre-processing activities and BoW approach as feature selection on classifying text

Table 3 The classification performances using Bag of Word approach

Experiment	No. of features	Accuracy	Precision	Recall	F-measure
1	198	79.17	76.47	85.25	80.62
2	126	77.75	75.76	81.97	78.74
3	122	86.67	80.00	100.00	88.89

Table 4 The classification performances using TF-IDF approach

Experiment	No. of features	Accuracy	Precision	Recall	F-measure
4	413	72.50	66.67	98.46	79.61
5	374	84.17	82.14	100	90.19
6	326	95.00	91.55	100	95.59

documents. Meanwhile, Experiments 4–6 are carried out to access the implication text pre-processing task and TF-IDF towards text classification performances. At the beginning, tokenization generates 935 features to be used for the entire experiments. However, duplicate or redundant features are removed to generate 562 unique features trough tokenization. Then, these features pool is used to reduce the number of features in stop-word removal task and stemming. In BoW approach, features with frequency of 2 and above are considered. For example, in Experiment 1, BoW only considers 198 features of 562 unique feature with frequency of 1. Later, it obtained 79.17% of accuracy and 80.62% of F-measure as shown in Table 3. Meanwhile, in Experiment 2 involving tokenization and stop-word removal, the classification performances is expected to be improved but the classification performances reduced. Even though stop-words are known as functional words, they also give a significant effect on the classification performance as removing stop-words decreases the classification performances. However, in Experiment 3 with tokenization, stop-word removal and stemming obtained the best classification improvement. It shows that the highest accuracy is achieved with the smaller number of features (122 features). In this case, stemming helps to reduce the number of dimensions as well as to carry out more unique features for classification by removing duplicate features (feature of having the same root word).

Meanwhile, for TF-IDF, referring to Table 4, the number of features obtained from each experiment produces more features compared to BoW approach. TF-IDF approach selects features by ranking features based on TF-IDF values and threshold. Thus, more features is considered, which hit better classification performances. In the Experiment 4, the classification accuracy is not promising but has a higher recall rate (98.46%), while the other two experiments achieved 100% recall rate. For accuracy, precision rate, and F-measure, they increase from Experiment 4 to 6 as the number of features decrease. It proved that features selected by TF-IDF are able to represent the feature's importance and relevance compared to BoW, which only selects features

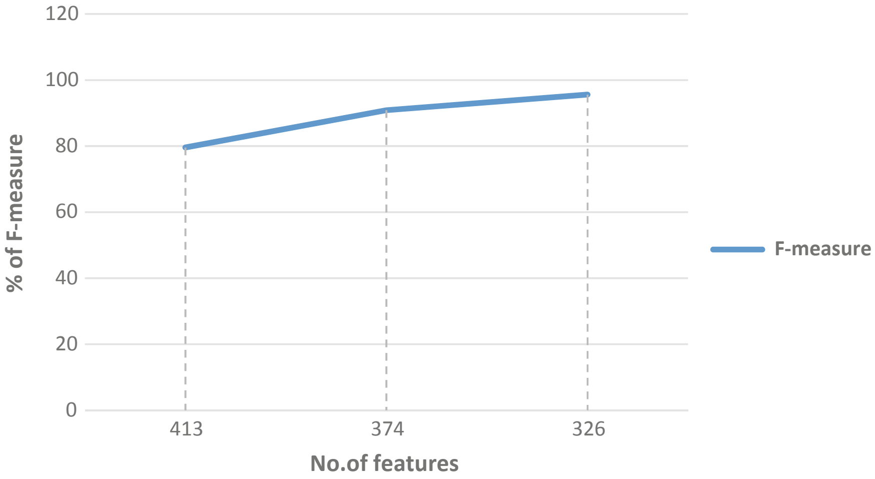


Fig. 8 Relationship between number of features and classification performances

based on the feature's frequencies. In addition, it reveals that the number of features has an impact on classification performances and the lower the number of features, the better the classification performances.

5 Conclusion

This paper addressed the effect of pre-processing on the performance of text classification. A tweet dataset is utilized to evaluate the effect of pre-processing activities and feature selection. Next, SVM classifier is employed to generate the final result. From the experimental result, it clearly revealed that pre-processing activities and feature selection gave a significant impact on the text classification performances. TF-IDF approach outperformed the traditional BoW approach. Figure 8 proved that there is a very significant relationship between the number of features and classification performances for TF-IDF feature selection with pre-processing activities.

Acknowledgements We would like to show our gratitude to Universiti Malaysia Pahang grant (RDU vote number RDU180380) for supporting this research.

References

1. Singh V, Saini B (2014) An effective pre-processing algorithm for information retrieval systems. Int J Database Manag Syst (IJDMS) 6(6):249–257
2. Syafiqah N, Nafis M, Awang S (2018) Challenges and issues in unstructured big data: a systematic literature review. Adv Sci Lett 24(10):7716–7722
3. Ceska Z, Fox C (2009) The influence of text pre-processing on plagiarism detection. In: RANLP, pp 55–59
4. Goncalves CA, Goncalves CT, Camacho R, Oliveira EC (2010) The impact of pre-processing on the classification of MEDLINE documents. In: Pattern recognition in information systems, proceedings of the 10th international workshop on pattern recognition in information systems, PRIS 2010, in conjunction with ICEIS 2010, p 10
5. Shinde AA, Chougule DG (2012) Text pre-processing and text segmentation for OCR. IJCSET 2(1):810–812
6. Jianqiang Z, Xiaolin GUI (2017) Comparison research on text pre-processing methods on Twitter sentiment analysis. IEEE Access 5:2870–2879
7. Al-Saffar A, Awang S, Tao H, Omar N, Al-Saiagh W, Al-bared M (2018) Malay sentiment analysis based on combined classification approaches and Senti-lexicon algorithm. PLoS ONE 13(4):1–18
8. Islam MS, Md Jubayer FE, Ahmed SI (2017) A support vector machine mixed with TF-IDF algorithm to categorize Bengali document. In: ECCE 2017—International Conference on Electrical, Computer and Communication Engineering, pp 191–196
9. Buckley C (1985) Implementation of the SMART information retrieval system, no. TR85-686, p 37
10. Porter MF (1980) An algorithm for suffix stripping
11. Bhaskar V (2017) Mining crisis information: a strategic approach for detection of people at risk through social media analysis. Int J Disaster Risk Reduct
12. Javed K, Maruf S, Babri HA (2015) A two-stage Markov blanket based feature selection algorithm for text classification. Neurocomputing 157:91–104

Developing and Validating an Instrument to Measure the Trust Effect Towards Instagram Sellers

Salwa Mustafa Din, Ramona Ramli and Asmidar Abu Bakar

Abstract Social commerce (s-commerce) has gained its popularity among users of social media sites as an online shopping platform including Instagram. Since its introduction in 2010, Instagram has become a chosen S-Commerce platform due to its easy to use features, its image attractiveness and its non-distracting interface. However, trust remains as the main issue in Instagram since Instagram involves mostly Consumer-to-Consumer (C2C) and transactions are made through private accounts. This study proposes an instrument to evaluate the trustworthiness of Instagram sellers and validates the items using statistical techniques. A questionnaire was designed and distributed as the instrument. Data was collected and analysed using SPSS to test on the item's reliability. 41 responses were collected for the pilot study, and the finding shows that the instrument is a valid and reliable measurement for the research model development.

Keywords Trust · Instagram · Purchase intention · s-commerce

1 Introduction

Several years ago, people were engaged in a new platform of commerce that was via social media platforms such as Facebook and Instagram. This new trend especially amongst millennials is called social commerce (s-commerce).

S. M. Din (✉)
College of Computing and Informatics (CCI), Universiti Tenaga Nasional (UNITEN),
Kajang, Malaysia
e-mail: salwamustafadin@gmail.com

R. Ramli · A. A. Bakar
Department of Computing, College of Computing and Informatics (CCI),
Universiti Tenaga Nasional (UNITEN), Kajang, Malaysia
e-mail: ramona@uniten.edu.my

A. A. Bakar
e-mail: Asmidar@uniten.edu.my

Z. Zakaria and R. Ahmad (eds.), *Advances in Electronics Engineering*,
Lecture Notes in Electrical Engineering 619,
https://doi.org/10.1007/978-981-15-1289-6_26

At the beginning, social media platforms were wholly intended for communication purposes and social interactions. However, as the platforms grow and advances with new tools, the platforms have become a social marketplace. The platforms can be seen as malls, where goods are being advertised and buying and selling activities are being made. At the same time, it remains as platforms to hang out and socialize as well.

Instagram is one of the platforms opted of s-commerce. It was first introduced in 2010, initially as a social platform for people to share pictures and images and socialize [1]. Recently, it has become a promising platform for buying and selling activities.

However, a study by Amelia (2016) found out that there were a lot of frauds happening on Instagram, an Indonesian case scenario. The study listed all possible frauds on both buyers and sellers' perspectives [2]. Frauds involving Instagram sellers such as 1. Product took longer than the due date with several reasons, 2. The product quantity and specification is different and is a defected product, 3. The seller claimed that the buyer has not sent the payment, 4. Blacklisting the buyer's account after payment, 5. Returning the wrong product and the seller do not want to pay the return fee, 6. Trusted seller at the first order to gain follower and customer, then cheating, 7. Hypnotize buyer to make a payment after the customer send their personal identity information, and 8. Fake discounts [2]. With frauds happening on the platform, thus, trust is seen vital to protect buyers on Instagram.

The study would identify factors influencing trust towards Instagram Sellers based on previous literatures, and proposes an instrument to measure trustworthiness of Instagram Sellers. A conceptual model is presented to show the relationship of factors with trust on Instagram Sellers. Then, the instruments would be validated using statistical techniques to ensure that the items used to measure the factors are reliable to be used in a survey.

1.1 Trust in Instagram

Since Instagram is chosen for s-commerce activities, both by buyers and sellers, trust is still an issue since Instagram is mostly Consumer-to-Consumer (C2C) and transactions are usually done using bank transfers to personal accounts. Sellers use the platform to advertise on products but buyers still have to message Instagram sellers for more information and for transaction purposes either sing WhatsApp or directly comment on the Instagram page. Trust in a C2C setting is vital since there is no security-based transaction method such as in e-commerce platforms [3].

Only a few studies were done on discussing factors that led to trust on Instagram [4–6]. Jasmine et al. (2017) proposed a trust framework from the viewpoints of social-psychological, sociological, and personality theories, and found that three factors were significant in influencing trust: Perceived Benevolence, Perceived Integrity, and Key Opinion Leadership [5]. Rafinda et al. (2018) studied on the variables that make Instagram online sellers trustworthy, from the customers' perspective and found that four factors proposed were significant in influencing trust: number of followers, Rationality of Price, image quality, and seller's response [6]. Gibreel et al. (2018) studied on how emerging markets influenced the way people shop online with Instagram as an s-commerce platform. The study showed that factors were interlocking and two factors were antecedents of trust in Instagram, which were familiarity and word of mouth [4]. From these previous literatures on trust on Instagram, a list of eight sub-factors were generated and proposed for the study.

2 The Conceptual Model

The study proposes a conceptual model to show the relationship of the factors with trust on Instagram sellers. From previous literatures on trust and purchase intention on Instagram, eight sub-factors were identified and grouped according to three Key Factors: (i) Seller-related factors, (ii) External factors, and (iii) Buyer-related factors). Table 1 shows the definition of each sub-factor adapted from previous research.

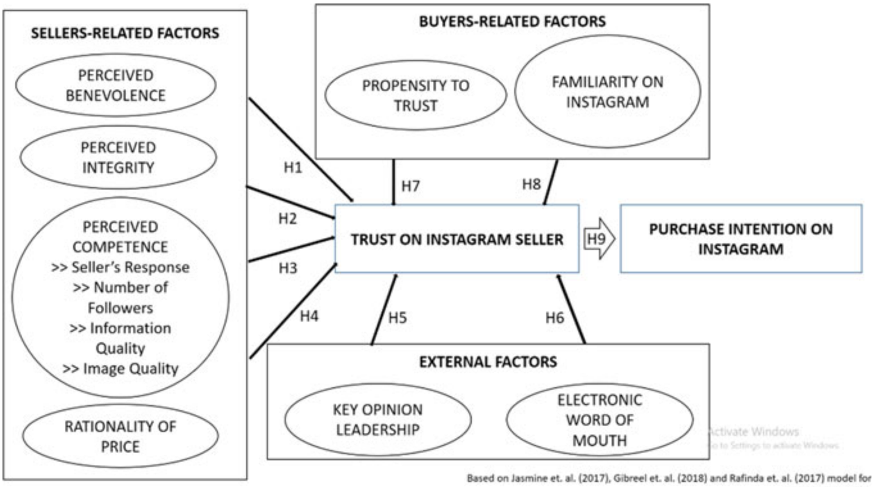
Table 1 Key factors and sub-factors for trust on Instagram

Key factors (sub-factors)	Definition	Authors
Sellers-related factors		
Perceived Benevolence (PB)	The perception that the Instagram sellers are providing care and good service to the buyers and potential buyers	Barnett White [7] Schoorman [8] Jasmine et al. [5]
Perceived Integrity (PI)	The perception that the Instagram seller shows consistency in actions and a fair buying-selling process to the buyers	Schoorman [8] Jasmine et al. [5]
Perceived Competence (PC)	The perception of the ability of the Instagram seller to display products with high quality photos, demonstrate knowledge on their products, show commitment through their speed and quality of response, and the number of accounts that follow the Instagram seller	Schoorman [8] Jasmine et al. [5] Rafinda et al. [6]
Rationality of Price (RP)	How fit the price tags of products are as compared to other Instagram sellers with regards to the quality of the products sold	Gibreel et al. [4] Rafinda et al. [6] Lichtenstein [9]

(continued)

Table 1 (continued)

Key factors (sub-factors)	Definition	Authors
External factors		
Key Opinion Leadership (KOL)	KOL are individuals who have established their authority on Instagram and are perceived as experts in specific areas, and their followers trust their recommendations	Egger [10] Jasmine et al. [5]
Electronic Word of Mouth (EWOM)	Statements on products or recommendations made by existing customers to potential customer	Gibreel et al. [4] Ramli et al. [11]
Buyers-related factors		
Familiarity (FM)	Being familiar with Internet marketplaces in the context of Instagram and having the knowledge in buying and selling process through Instagram	Gibreel et al. [4] Zadmehr et al. [12]
Propensity to Trust (PTT)	The psychological perspective focusing on a personality trait that is willing to depend and trust others	Schoorman [8] Jasmine et al. [5] Zadmehr et al. [12]



Based on Jasmine et. al. (2017), Gibreel et. al. (2018) and Rafinda et. al. (2017) model for trust on Instagram

Fig. 1 Conceptual model for trust on Instagram seller

Figure 1 shows the proposed conceptual model based on the generated identified factors. Eight sub-factors are identified as independent variables for trust on Instagram sellers: Perceived Benevolence (PB), Perceived Integrity (PI), Perceived

Competence (PC), (RP), Key Opinion Leadership (KOL), Electronic Word of Mouth (EWOM), Propensity to Trust (PTT), and Familiarity on Instagram (FM). Buyers' trust on Instagram sellers would be determining purchase intention (PUI) on Instagram.

3 Research Methodology

3.1 Instrument Development

A questionnaire was constructed and adopted from previous literatures [4–6, 8, 9, 11, 12]. The answers for the items ranged on a 5-point Likert scale (Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree). Finally, respondents were asked to answer some questions related to purchase intention, also on a 5-point Likert scale. The study by Melanie (2014) shows that if researchers want to use agree-disagree scales, they should offer a 5 answer categories rather than 7 or 11, because the latter yield data of lower quality [13]. The online pilot survey consisted of four (4) sections: (i) general information on respondents, (ii) previous online experience and length of using Instagram, (iii) information on the Instagram seller, and (iv) items which would assess all of the eight sub-factors.

3.2 Pilot Survey

The pilot survey was distributed via e-mail and links were given to respondents using WhatsApp messages. A total of 41 respondents, who were postgraduate students and researchers, were involved in the pilot survey. An analysis of reliability of items were assessed using SPSS software, and Cronbach alpha values were calculated in order to assess the reliability of items for each factor assessed. According to Gliem and Gliem (2003), it is imperative to calculate and report on the Cronbach's alpha coefficient values for assessing internal consistency reliability when using Likert-type scales [14]. Table 2 shows the results of the reliability analysis for each sub-factor assessed. The initial reliability analysis showed that all of the eight sub-factors had Cronbach alpha values of a range between 0.667 and 0.930. Two sub-factors were reviewed in terms of deleting some items: (i) Perceived Competence (2 items from the number of followers), and (ii) Rationality of Price (1 item). The rest of the sub-factor items were retained for the real survey.

The value of Cronbach alpha for Perceived Competence increased from the initial value of 0.850–0.863 after removing two items which are: (*a*) *I believe that the higher the number of followers of the Instagram seller, the more likely that the seller can be trusted* and (*b*) *I believe that the lower the number of followers of the*

Table 2 Cronbach Alpha values for items on each sub-factor

Sub-factors	Number of items	Cronbach Alpha value	Cronbach Alpha value (after deleting item(s))	Number of items (after deleting item (s))
Perceived Benevolence (PB)	3	0.847	–	3
Perceived Integrity (PI)	3	0.917	–	3
Perceived Competence (PC)	12	0.850	0.863	10
Rationality of Price (RP)	3	0.667	0.726	2
Key Opinion Leadership (KOL)	3	0.930	–	3
Electronic Word of Mouth (EWOM)	8	0.893	–	8
Familiarity (FM)	3	0.862	–	3
Propensity to Trust (PTT)	4	0.767	–	4
Purchase Intention (PUI)	3	0.916	–	3

Instagram seller, there is an insecure feeling on the seller. The value for Cronbach alpha for Rationality of Price increased from 0.667 to 0.726 after removing the item: (*a*) *I am worried about the fraud if there are goods with low quality but have a high price tag*. Therefore, the internal consistency of the items after deleting some bad items are considered accepted to very good [14].

3.3 Data Collection and Analysis

The pilot survey was distributed to 41 Postgraduate students and researchers who have Instagram accounts and have previous experience in buying online. The respondents were asked to recall previous experience with an Instagram seller and were then asked to answer an online survey.

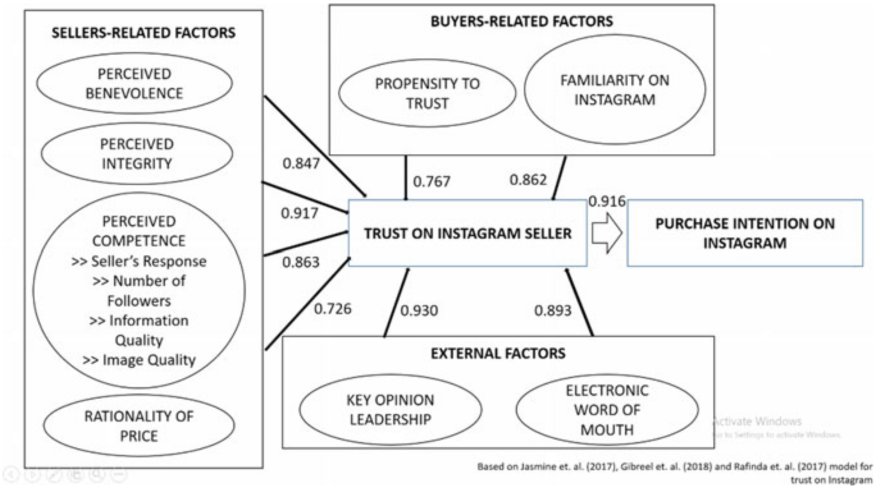
The respondents of this study were 75.6% female and 24.4% male. The majority of respondents were 25–34 years old representing half of the respondents. 78% of the respondents have been using Instagram for more than 2 years. Majority of the respondents were Malay with 85%. The Instagram seller being assessed the most was having between 1000 and 9999 followers on Instagram, and the highest category of products sold by the Instagram seller is in the category of Clothing, Accessories, and Shoes. The following Table 3 summarizes the respondents' demographics.

Table 3 Key factors and sub-factors for trust on Instagram

Measures	Item	Percentage (%)
General information respondents		
Gender	Male	76
	Female	24
Age	24 and below	19
	24–34	51
	35–44	20
	45 and above	10
Race	Malay	85
	Chinese	2
	Indian	5
	Others	7
Monthly Income	Below RM1000	15
	RM1000–RM2999	39
	RM3000–RM4999	12
	RM5000 and above	34
Length of time using Instagram	<6 months	7
	6 months–1 year	10
	1 year–2 years	5
	>2 years	78
Information on Instagram seller		
Number of Followers	<1000 Followers	24
	1000–9999 Followers	34
	10,000–99, 999 Followers	22
	100,000–999,999 Followers	10
	>1,000,000 Followers	10
Category of Products sold	Cosmetics, Skincares, Health Supplements	7
	Books, Electronic Materials	7
	Clothing, Accessories, Shoes	39
	Toys	0
	Computer, Electrical & Communication Appliances	7
	Combination of more than one (1) category	37

4 Data Analysis and Results

The internal consistency values for the sub-factors were: PB = 0.847, PI = 0.917, PC = 0.863, RP = 0.726, KOL = 0.930, EWoM = 0.893, FM = 0.862, PTT = 0.767. The purchase intention item instruments' Cronbach Alpha value was PUI = 0.916. Figure 2 shows the Cronbach alpha values for the eight sub-factors,



Based on Jasmine et. al. (2017), Gibreel et. al. (2018) and Rafinda et. al. (2017) model for trust on Instagram

Fig. 2 Cronbach alpha values of each sub-factor

which influences trust, and purchase intention on Instagram. Since all Cronbach alpha values were more than 0.7 (considered good to very good), then the instruments used to assess the sub-factors would be considered to be used for the real survey.

5 Conclusion

This study proposed an instrument to measure the trustworthiness of Instagram sellers. For this purpose, a questionnaire was constructed and adopted from previous literatures on trust in Instagram. A conceptual model is also proposed to show the factors as antecedents and independent variables of trust, adopted from three different models by [4–6]. Reliability analysis was done to assess the internal consistencies of the eight sub-factors towards the trustworthiness of an Instagram seller. The developed instruments were also validated using simple statistics. Results showed that after deleting three items, all Cronbach Alpha values showed good to very good internal consistency values, hence, indicating that the instrument is reliable for the proposed survey. The proposed conceptual model would be implemented in future studies, and data collection can be done by involving Instagram users in Malaysia. The model would be evaluating the trustworthiness of Instagram sellers, which would influence the purchase intention on Instagram.

Acknowledgements This research was supported by the Ministry of Higher Education, Malaysia under FRGS/1/2015/SS01/UNITEN/03/2.

References

1. Tech Crunch Site. https://techcrunch.com/2012/04/24/the-rise-of-instagram-tracking-the-apps-spread-worldwide/. Last accessed 15 May 2019
2. Amelia TN (2015) Fraud in online transaction: case of Instagram. J Adv Manage Sci 4 (4):347–350
3. Jones K, Leonard LNK (2008) Trust in consumer-to-consumer electronic commerce. Inf Manage 45(2):88–95
4. Gibreel O, AlOtaibi DA, Altmann J (2018) Social commerce development in emerging markets. Electron Commer Res Appl 27:152–162
5. Che JWS, Cheung CMK, Thadani DR (2017) Consumer purchase decision in Instagram stores: the role of consumer trust
6. Rafinda A, Suroso A, Purwaningtyas P (2018) Formative variables of trustworthiness on Instagram online sellers, vol 25, no 2, pp 1–7
7. Barnett White T (2005) Consumer trust and advice acceptance: the moderating roles of benevolence, expertise, and negative emotions. J Consum Psychol 15(2):141–148
8. Mayer RC, Davis JH, Schoorman FD (2014) An integrative model of organizational trust. Acad Manage Rev 20(3):709–734 (Pearson)
9. Lichtenstein DR, Ridgway NM, Netemeyer RG (2006) Price perceptions and consumer shopping behavior: a field study. J Mark Res 30(2):234
10. Egger FN (2003) Trust me, I'm an online vendor, p 101
11. Ramli R, Abu Bakar A, Aziz N (2016) Review of literatures on trust factors affecting customer's purchase intention on SOHO seller: S-commerce context. Inf J 19(7B):2791–2796
12. Zadmehr S, Fatourehchi S, Zadmehr H (2016) Identifying and structuring factors affecting the purchase from mobile social networks (Case Study: Instagram Application). Int J Life Sci 10(2)
13. Revilla MA, Saris WE, Krosnick JA (2014) Choosing the number of categories in agree-disagree scales. Sociol Methods Res 43(1):73–97 (Sage Publications)
14. Gliem RR, Gliem JA (2003) Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for likert-type scales. In: 2003 Ohio State University Conference, Columbus, Ohio

Technical Feasibility and Econometrics of a Small Hydro Power Plant for a Farm Settlement in Nigeria

Felix A. Ishola, Ademola Dare, Joseph Azeta, Timilehin Sanni and Stephen A. Akinlabi

Abstract The farm settlement comprises of large hectares of crop plantation and a community of less than a hundred number of people living in low-cost buildings. The farm settlement project is a pilot public-private project majorly to facilitate crop production and encourage urban-rural settlement for the purpose of boosting crop exportation capacity of the country. There had been a difficulty in electrifying the farm settlement using the existing national grid. The nearest power station is about 1000 km from the farmland and the energy capacity available from the power booster station has already been chocked up by some closer and bigger communities. This study explored a sustainable energy source for a regular and dependable power supply for the farm settlement. The authors investigated the feasibility of having an electric power generation from a local river close to the farm settlement. Also, the econometrics for the Small Hydro Power Scheme was carried out to determine the viability of the proposed project.

Keywords Econometrics · Feasibility · Hydro Power Plant · Settlement

F. A. Ishola (✉) · J. Azeta · S. A. Akinlabi
Mechanical Engineering Department, Covenant University, Ota, Nigeria
e-mail: felix.ishola@covenantuniversity.edu.ng

A. Dare
Mechanical Engineering Department, University of Ibadan, Ibadan, Nigeria

T. Sanni
Electrical and Electronic Engineering Department, Covenant University, Ota, Nigeria

S. A. Akinlabi
Department of Mechanical Engineering, Walter Sisulu University,
East London 5200, South Africa

Z. Zakaria and R. Ahmad (eds.), *Advances in Electronics Engineering*,
Lecture Notes in Electrical Engineering 619,
https://doi.org/10.1007/978-981-15-1289-6_27

1 Introduction

The farm settlement project is a public-private project established to facilitate crop production thus boosting crop exportation capacity of the country. There had been a difficulty in electrifying the farm settlement using the existing national grid. The nearest power station is about 1000 km from the farmland and the energy capacity available from the power booster station has already been chocked up by some closer and bigger communities. A survey of the river course was undertaken to assess the hydropower potential of the river. The river is generally fed from runoff from surrounding hills which are about 650 ft. The runoff formed the different size of gulge in the pathway of the river. An assessment of the different sizes of gulge revealed that at some location along the river course, with a suitable weir, a gross head of about 15 m could be achieved. This was therefore considered encouraging for a hydropower plant [1–3].

1.1 A Sub Topographical, Perimeter and Hydro-Meteorological Survey of the Project Location

Please The project area falls within the basement complex rock formations and these rocks are found to be rich of quartzites and amphibolite which form part of the southern segment of the Schist belt of Nigeria and are relatively impermeable. Although hilly rock outcrops are scarce except for quartzites rubbles which can be found mainly along the side of rivers. Even in weathered conditions, the rocks break down to clays and soil types which in their compositions still maintain the relative impermeability. The topsoil in this area is generally of porous, friable, moderately drain profiles. Below this layer is the organic clay fraction of relatively low permeability. Except in marked basins, the water table is quite low even at the peak of the rainy season. The project area is characterized by two distinct seasons, a dry season which lasts for about eight months begins in March and lasts until October. While the wet season lasts for about four months, from November to February. For the project area, the lowest daily minimum temperatures are experienced during the wet season but sometimes very low values occur in December and January due to the cold harmattan breeze. The average daily highest temperature is observed during the dry seasons starting from November with about 27.20 °C, coming to a peak in April at about 28.30 °C, and falling to the lowest of 25.00 °C in August [4].

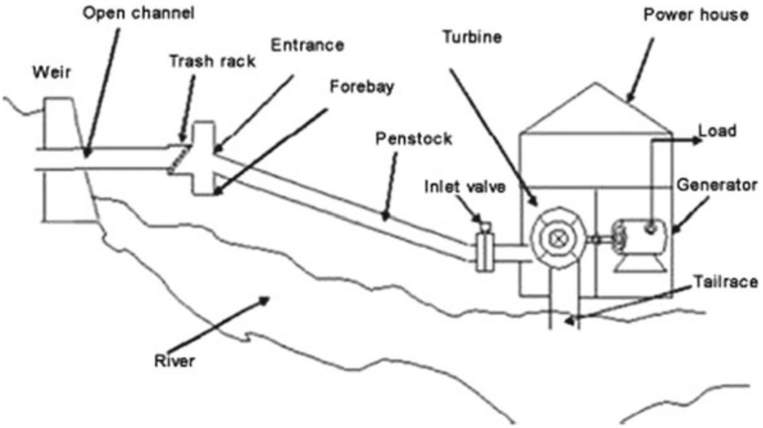


Fig. 1 Schematic view of the proposed plant [7, 8]

2 Proposed Hydropower Plant

From the geological and topographic studies, the river course has a relatively flat terrain. However, since there are deep gulge along the river course a small dam can be constructed to gain an appreciable head for a suitable hydropower plant. Based on this, the proposed hydropower scheme is shown in Fig. 1. In operation, the penstock which is expected to be about 100 m in length discharges water at design flow of 20 m^3/s into the turbine and subsequently drive the generator via speed enhancer and thus generating the required electrical power [1, 5, 6].

2.1 Location of the Dam and Power House

Since there are deep gullies along the river course, the dam is proposed to be located at a point where there is at least 20 m deep gully. The powerhouse is to be located relative to the dam axis ensuring that the gross head is not compromised [9].

3 Determination of Suitable Gross Head for the Hydropower Plant

In determining a suitable head for the hydropower plant, a financial assessment was carried out. Generally, small hydropower plant is considered viable if the payback is about 7 years. Some gross heads were then considered. The results are shown in Table 1. While a smaller head of about 5 m would have saved cost, the payback is

Table 1 Cost analysis table for dam height selection

Dam height (m)	Investment cost (N)	Payback (years)
5	652,000,000	14
10	722,000,000	7
15	815,000,000	6

on the high side. Higher head such as 15 m would have produced higher power output, but the cost seems unfavourable. On the final analysis, a gross head of 10 m was found to be financially viable [8–11].

3.1 Detailed Description of the Scope of Works

The Project involves the construction of a hydropower plant via the use of a reservoir.

3.2 The Power House

The Power House: a one-room structure located directly in the downstream section of the fill dam body, directly adjacent to the left-wing wall of the spillway structure. The powerhouse comprises of the generating unit and all control and auxiliary equipment [12, 13]. The powerhouse shall be equipped with a crane having a lift capacity of 40 kN.

3.3 Mechanical Equipment

The major design considerations for mechanical equipment are as follows:

- High availability and reliability of the plant,
- Operation close to optimum efficiencies,
- Reliability and Quality of equipment parts, components and materials,
- Easy maintenance and repair of equipment [14].

3.4 Hydraulic Parameters for Designing the Turbine

The hydraulic conditions needed for the selection of the turbine are 10 m Gross hydraulic head, 9 m Net hydraulic head and 20 m^3/s. The selection of the optimum turbine size, specific and synchronous speed has been done by comparison of various alternatives [10, 15–18]. Thus, one horizontal Francis Turbine will be installed in the power plant. The Turbine connects directly to the generator shaft. The turbine runner and guide vanes shall be made of stainless cast steel and thus highly resistant to cavitations and silt erosion [10, 19–21]. The design of the turbine shall be such that it allows the dismantling of the runner and guide vanes mechanism without affecting the generator arrangement [22].

To prevent pollution of the water by lubricants and to facilitate the maintenance works, all guide vanes bearings and joints for regulating mechanism are the self-lubricating types. The turbine is equipped with a butterfly valve to shut-off the turbine against the headwater. The butterfly valve will be closed under the no-flow condition, after closing the guide vanes of the turbine. The valve will also serve as an emergency shut-off device in case the guide vanes fail to close, as well as, failure in power supply and pressure oil supply [23].

3.5 Hydraulic Control Module

The turbine will be equipped with a hydraulic control module for the operation of the Turbine Guide Vanes and the Main Inlet Valve. The Hydraulic Control Module will also ensure a stable governing in parallel operation and also in the isolated operation of the power plant [18].

3.6 Powerhouse Crane

One electrically driven, single girder, single trolley floor controlled, indoor type bridge crane of 40 kN capacity will be used primarily for unloading the components of the electromechanical equipment and for the installation [24].

3.7 Electrical Equipment

The purpose of the electrical design studies was to identify the main components of the electrical equipment in and around the powerhouse for safe and economic operation of the hydropower plant. The electrical equipment consists mainly of the following:

- One Generator, 1000 kVA, 600 rpm, 50 Hz, 0.4 kV
- 1 main step-up Transformer 1000 kVA, 0.4 kV/11 kV
- MV switchgear and overhead line
- Low voltage switchgear

- DC system
- Cabling, lighting system and small power installations
- Earthing and lightning protection systems
- Control, protection and communication systems.

Both the generator and transformer are expected to be upgraded at the later stage of project life [24].

3.8 Generator

The generator will be of 3-phase, synchronous, horizontal type directly coupled to the Francis Turbine. A flywheel will be arranged between the Turbine and the Generator. The features of the Generator are governed mainly by the results of the mechanical engineering design studies [25].

3.9 Main Step-up Transformer

To raise the generating voltage to the medium voltage distribution level, a main step-up transformer will be provided. The transformer will be installed outdoors at the transmission line terminal pole [26].

3.10 Plant Auxiliary Systems

The LV switchgear will be of metal clad. Indoor type, factory assembled and type tested, comprising fixed amount equipment for main and auxiliary circuits [27]. To supply the control, protection and communication systems, a 24 V DC system will be provided consisting 1unit quantity of maintenance free, gas-tight lead acid battery; 1unit quantity of battery charger 400 V AC/24 V Dc and 1unit quantity of DC distribution panel.

3.11 Distribution Line 11 kV

The distribution system in that region consists mainly of 11 kV lines. To transmit power from the Hydropower plant, which is in the range of 1100 kVA, the distribution voltage level 11 kV is selected. The hydropower plant connects to the existing 11 kV systems, feeder Overhead Transmission line existing at the project site.

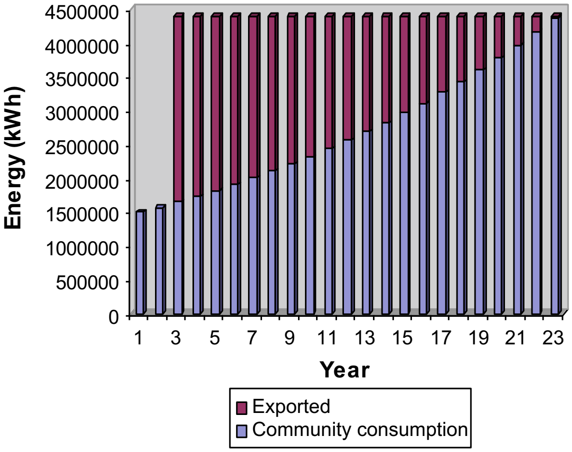


Fig. 2 Electricity consumption forecast for the farm settlement

4 Electricity Consumption Model Energy Generation and Evacuation Synergy

The energy consumption of the community is projected to gradually increase from the present estimate of 1,500,000–4,376,000 kWh. As such some of the energy produced can be exported via connection to the national grid. A graphical rendition of energy exported as compared with the community consumption is shown in Fig. 2.

4.1 Financial Analysis of the Proposed Hydropower Plant

The viability of the entire project depends on the accrued revenue and as well the project costs. These are hereby discussed.

4.2 Forecasted Revenue

The energy generated will be utilized by the target project area while the rest will be exported via the national grid. Using N8/kWh as the average the cost of electricity in Nigeria and based on this the projected revenue within the life span of the project is presented in Table 2.

Table 2 Anticipated revenue from the operation of the hydropower plant

S/N	Year	Estimated energy consumption/day	Revenue for the year (load centre) (N)	Revenue from export (N)	Total revenue (N)
1	2019	839,500			
2	2020	881,475			
3	2021	925,464	7,403,712	21,984,000	29,387,712
4	2022	971,469	7,771,752	21,344,000	29,115,752
5	2023	1,019,824	8,158,592	20,640,000	28,798,592
6	2024	1,070,530	8,564,240	19,904,000	28,468,240
7	2025	1,123,922	8,991,376	19,136,000	28,127,376
8	2026	1,180,001	9,440,008	18,336,000	27,776,008
9	2027	1,238,766	9,910,128	17,504,000	27,414,128
10	2028	1,300,553	10,404,424	16,640,000	27,044,424
11	2029	1,365,362	10,922,896	15,712,000	26,634,896
12	2030	1,433,530	11,468,240	14,720,000	26,188,240
13	2031	1,505,055	12,040,440	13,696,000	25,736,440
14	2032	1,580,274	12,642,192	12,640,000	25,282,192
15	2033	1,659,187	13,273,496	11,488,000	24,761,496
16	2034	1,742,130	13,937,040	10,304,000	24,241,040
17	2035	1,829,102	14,632,816	9,056,000	23,688,816
18	2036	1,920,440	15,363,520	7,776,000	23,139,520
19	2037	2,016,143	16,129,144	6,400,000	22,529,144
20	2038	2,116,883	16,935,064	4,960,000	21,895,064
21	2039	2,222,660	17,781,280	3,456,000	21,237,280
22	2040	2,333,474	18,667,792	1,856,000	20,523,792
23	2041	2,449,996	19,599,968	192,000	19,791,968
Total			264,038,120	267,744,000	531,782,120

4.3 Project Costs

The project cost highlighted the cost for the implementation of the envisaged design for the hydropower plant. It should be noted that the miscellaneous cost has been put at about 7% of the project cost. Table 3 shows a summary of the project cost.

4.4 Operational and Maintenance Cost

These include the cost of labour for the operation and maintenance of the hydropower plant. Generally, a small hydropower plant does not require much of the operator's attention and barely need a high maintenance scheme. An estimate of

Table 3 Summary of the project cost

Initial costs (credits)	NGN
Power system	
Hydro turbine	414,957,000
Road construction (to the plant)	22,769,000
Transmission line	12,656,000
Substation	2,995,000
The balance of system & miscellaneous	
Penstock	64,292,000
Other	201,433,000
Total initial costs	**719,101,000**

N3m annually is estimated to be sufficient enough for the smooth operation of the plant. This hopefully can be used to fully engage an operator and as well meet up with other maintenance exigencies.

4.5 *Cash Flow Analysis*

A financial analysis was carried out using the projected income and envisaged total costs of the project. A cash flow table generated is presented in Table 4. This is again rendered in graphical form as shown in Fig. 3. A prevailing inflation rate of

Table 4 Expected cash flow for the project

Year	Pre-tax	After-tax	Cumulative
#	NGN	NGN	NGN
0	−700,000,000	−700,000,000	−700,000,000
1	88,650,383	88,650,383	−611,349,617
2	88,472,935	88,472,935	−522,876,681
3	88,280,937	88,280,937	−434,595,745
4	88,073,194	88,073,194	−346,522,551
5	87,848,416	87,848,416	−258,674,135
6	87,605,207	83,224,947	−175,449,188
7	87,342,055	82,974,952	−92,474,236
8	87,057,324	82,704,458	−9,769,778
9	86,749,245	82,411,783	72,642,004
10	86,415,903	82,095,108	154,737,113
11	86,055,228	81,752,467	236,489,579
12	85,664,977	81,381,729	317,871,308
13	85,242,726	80,980,590	398,851,898

(continued)

Table 4 (continued)

Year	Pre-tax	After-tax	Cumulative
14	84,785,850	80,546,558	479,398,455
15	84,291,511	80,076,935	559,475,390
16	83,756,635	79,568,803	639,044,194
17	83,177,900	79,019,005	718,063,198
18	82,551,708	78,424,123	796,487,321
19	81,874,169	77,780,460	874,267,781
20	81,141,071	77,084,017	951,351,798

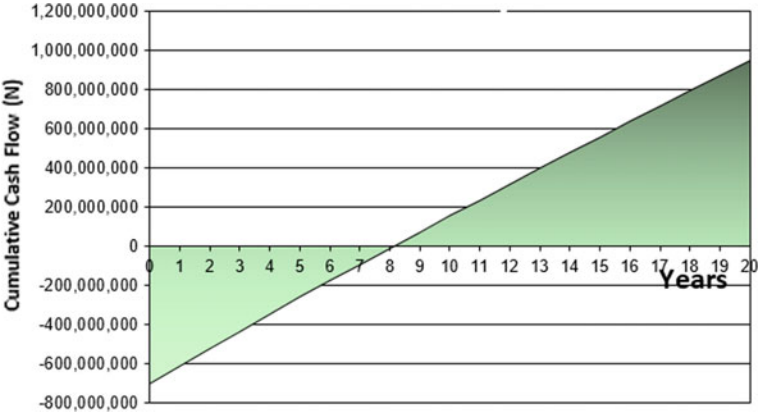


Fig. 3 The expected cash flow showing breakeven at the 8th year

8.2% was used for the analysis. A tax holiday of 5 years was assumed for the project while a depreciation rate of 5% was incorporated. From the result, it could be observed that the project will be able to generate substantial profit after eight years of operation [15, 28]. This projection could be improved upon by having a slight increase in electricity pricing.

5 Conclusion

This study explored a sustainable energy source for a regular and dependable power supply for the farm settlement. The authors investigated the feasibility of having an electric power generation from a local river close to the farm settlement. The econometric analysis for the Small Hydro Power Scheme was carried out and it was discovered that the Small Hydro turbine power generation project is viable with a payback in eight (8) years after which bountiful profit is projected.

References

1. Bitar Z, Khamis I, Alsaka Z, Al Jabi S (2015) Pre-feasibility study for construction of mini hydro power plant. Energy Proc 74:404–413
2. Jawahar CP, Michael PA (2016) A review on turbines for micro hydro power plant. Renew Sustain Energy Rev 72, no. October 2015, pp 882–887, 2017. Author F (2016) Article title. Journal 2(5):99–110
3. Mite-león M, Barzola-monteses J (2018) Statistical model for the forecast of hydropower production in Ecuador, vol 8, no 2
4. Signe K, Bertrand E, Hamandjoda O, Antoine FN, Takam G, Bertrand C (2017) Modeling of rainfall-runoff by artificial neural network for micro hydro power plant: a case study in Cameroon, pp 15511–15519
5. Loots I, Van Dijk M, Barta B, Van Vuuren SJ, Bhagwan JN (2015) A review of low head hydropower technologies and applications in a South African context. Renew Sustain Energy Rev 50:1254–1268
6. Gaiser K, Erickson P, Stroeve P, Delplanque JP (2016) An experimental investigation of design parameters for pico-hydro Turgo turbines using a response surface methodology. Renew Energy 85:406–418
7. Signe EBK, Hamandjoda O, Nganhou J, Wegang L (2017) Technical and economic feasibility studies of a micro hydropower plant in Cameroon for a sustainable development. J Power Energy Eng 05(09):64–73
8. Signe EBK, Hamandjoda O, Nganhou J (2017) Methodology of feasibility studies of micro-hydro power plants in Cameroon: case of the micro-hydro of KEMKEN. Energy Proc 119:17–28
9. Elbatran AH, Yaakob OB, Ahmed YM, Shabara HM (2015) Operation, performance and economic analysis of low head micro-hydropower turbines for rural and remote areas: a review. Renew Sustain Energy Rev 43:40–50
10. Sangal S, Garg A, Kumar D (2013) Review of optimal selection of turbines for hydroelectric projects. Int J Emerg Technol Adv Eng 3(3):424–430
11. Pratap Nair M, Nithiyananthan K (2016) Feasibility analysis model for mini hydropower plant in Tioman Island, Malaysia. Distrib Gener Altern Energy J 31(2):36–54
12. Cazzago D (2013) Technical and economical feasibility of micro hydropower plants
13. Singh VK, Singal SK (2017) Operation of hydro power plants-a review. Renew Sustain Energy Rev 69:610–619
14. Ho-Yan B (2012) Design of a low head pico hydro turbine for rural electrification in Cameroon
15. Zhou D, Deng ZD (2017) Ultra-low-head hydroelectric technology: a review, vol 78, pp 23–30
16. Ebhota WS, Inambao F (2016) Design basics of a small hydro turbine plant for capacity building in sub-Saharan Africa. Afr J Sci Technol Innov Dev 8(1):111–120
17. Nasir BA (2014) Design considerations of micro-hydro-electric power plant. Energy Proc 50:19–29
18. Williamson SJ, Stark BH, Booker JD (2014) Low head pico hydro turbine selection using a multi-criteria analysis. Renew Energy 61:43–50
19. Prajapati PVM, Patel PRH, Thakkar PKH (2015) Design, modeling & analysis of Pelton wheel turbine blade, vol 3, no 10, pp 159–163
20. Nigussie T, Engeda A, Dribssa E (2017) Design, modeling, and CFD analysis of a micro hydro Pelton turbine runner: for the case of selected site in Ethiopia. Int J Rotating Mach
21. Ishola FA, Azeta J, Agbi G, Olatunji OO, Oyawale F (2019) Simulation for material selection for a Pico Pelton turbine's wheel and buckets. Proc Manuf 30
22. Zainuddin H, Yahaya MS, Lazi JM, Basar MFM, Ibrahim Z (2009) Design and development of Pico-hydro generation system for energy storage using consuming water distributed to houses. Int J Electr Comput Energ Electron Commun Eng 3(11):154–159

23. Khomsah A, Sudjito, Wijono, Laksono AS (2019) Pico-hydro as a renewable energy: local natural resources and equipment availability in efforts to generate electricity. In: IOP conference series: materials science and engineering, vol 462, p 012047
24. AHEC (2012) Standards/Manuals/Guidelines for small hydro development
25. Kirmani S, Jamil M, Akhtar Y (2017) Bi-directional power control mechanism for a microgrid hybrid energy system with power quality enhancement capabilities, vol 7, no 4
26. Breeze P (2019) Power generation technologies, 3rd edn. Elsevier
27. Oyedepo SO, Fagbenle RO (2011) A study of implementation of preventive maintenance programme in Nigeria power industry—Egbin thermal. In: Energy power engineering, vol 2011, July, pp 207–220
28. Ajayi OO, Ohijeagbon OD (2017) Feasibility and techno-economic assessment of stand-alone and hybrid RE for rural electrification in selected sites of South Eastern Nigeria. Int J Ambient Energy 38(1):55–68

The Technological Implications of Using MEMS-Based Generators as a Replacement to Batteries in Electronic Devices—A Short Review

Timilehin Sanni, Felix A. Ishola, Olumide A. Towoju and Esther T. Akinlabi

Abstract Batteries had been widely used as the power source for remote electronic devices, but the method of generating energy from the environment had been with a lot of prospects. To ensure constant power supply and avoid complexity of changing batteries for wireless electronics and sensors, especially in the case of usages in remote areas, self-energy generating devices have become a necessity. The authors briefly review alternative to batteries; options such as electrostatic generation, dielectric elastomers and piezoelectric materials which are considered very promising for their capacity to change stains in the material into electrical vitality and be incorporated into electronic gadgets. The paradigm shift in the technological world towards producing electronic devices with focus on reduction in size, cost and energy consumption have constantly considered the generator based on Micro-electro-mechanical systems (MEMS). This article highlights the impacts and challenges of using MEMs based generators for providing power sources in small wireless sensor nodes, in place of batteries.

Keywords MEMS · Remote energy · Batteries · Wireless sensors

T. Sanni
Electrical and Electronic Engineering Department, Covenant University, Ota, Nigeria

F. A. Ishola (✉) · E. T. Akinlabi
Mechanical Engineering Department, Covenant University, Ota, Nigeria
e-mail: felix.ishola@covenantuniversity.edu.ng

O. A. Towoju
Mechanical Engineering Department, Adeleke University, Ede, Nigeria

E. T. Akinlabi
Department of Mechanical Engineering Science, University of Johannesburg, Johannesburg 2006, South Africa

Z. Zakaria and R. Ahmad (eds.), *Advances in Electronics Engineering*,
Lecture Notes in Electrical Engineering 619,
https://doi.org/10.1007/978-981-15-1289-6_28

1 Introduction

Micro-electro-mechanical systems (MEMS) are small devices that function as an integrated circuit (IC) and interrelating with their environments [1]. They are usually used in devices with low frequencies, low power consumption that dissipates at a relatively faster rate and at a lower cost. These are devices whose length is less than 1 mm, but more than 1 micron [2]. The method of manufacturing combines electronic and mechanical systems as with manufacturing IC components [2]. This area had been considered to be continuously growing due to a law postulated by Moore; the law implies that the use of transistors is increasing with the production of small systems. Silicon had been found to be the major raw material for the production of MEMS and NEMS, due to the fact that they are less expensive (taking into consideration that they are obtainable from the high purity class of quartz sand) and they fit into electro-mechanical systems easily [3]. Microsystem manufacturing requires high purity and reliability for efficient production. Silicon has proven its reliability. There are also already manufacturing processes for silicon.

A MEMS-based piezoelectric had been striving on the fundamental principle of using micro-structured cantilever beam for generating vibration using piezoelectric materials made from ceramic; PZT (lead zirconate titanate, Pb(Zr, Ti) $O_{3)}$ or semiconductors that can convert vibration into electro-energy [4, 5]. Two functional materials are used: Si and PZT, Si because it’s density is high and PZT due to its coefficient of electromechanical coupling [5]. This is as shown in Table 1. Piezoelectric generation becomes a favourite method because it is compatible with MEMS and there is no need for an external power source. It has sensitivity to low vibrations and low frequencies coming from surrounding [4].

1.1 Principles of Operation

A Cantilever beam is basically any plane structure that utilizes one open end to bear the load with supports at the opposite end. It has mechanical oscillations as inertia energy at a low frequency and kinematics can be manoeuvred into electrical energy [6]. The energy created from the environment thus can be utilized for electronic gadgets with low power consumption making use of MEMS fabrication process. With the use of MEM, a self-controlled sensor gadget can be utilized for the

Table 1 Properties of silicon material and PZT which makes them suitable for piezoelectric-generation [5]

Material	Density (kg/μm^3)	Modulus of elasticity (MPa)	Poisson’s ratio
PZT	7.55e−15	8.0e+4	0.25
Silicon	2.5e−15	1.69e+5	0.3

creation of low power utilization, in which 1.15 μW were delivered [7]. Converter utilizing the standard CMOS process is utilized to control a computerized system that requires 18 μW [8]. The generator comprises of silicon material dependent on CMOS and cantilever, which comprises various dimensions of material [7]. The vibration of nature and the dynamic energy is converted into electrical energy by the piezoelectric-impact.

Generation of strain occurs on the layered cantilever at a slightest vibration from the environment thus extending multi-layer cantilever movement. The input energy causes a load on piezoelectric layer through signals converted by the deflection of the beam which then produces a strain that generates electric charges [4]. The upper and lower cathodes serve as a storage for the charges generated. The voltage between the two terminals of the PZT layer decides the magnitude of the voltage. Maximum Energy transformation happens when the vibration of the cantilever beam is equal to the frequency of the input [4], It is vital that the input energy frequency compares to the frequency with which the test mass sways [9], with the outcome, that the most extreme power is produced with the best deviation. Consequently, the mass decides the voltage of the PZT while the electrical power relies upon the cantilever beam's shape and size as well as the mass of silicon. This is schematically represented in Fig. 1.

The power produced by the movement of the silicon mass builds maximum separation limit which the proof mass can go. The yield power is relative to the mass of silicon and furthermore relies upon the transformation of electrical and mechanical energy and the damping system [4]. Piezoelectric material is considered to be a creator of electrical power, because of its capacity to convert mechanical energy into electrical energy [9, 11]. Piezoelectric materials deliver energy anytime force is applied to dis-stabilize the gravity in-between the negative and positive charges of an atom [12]. They can withstand loads and convert them into electrical energy. They produce charges under load or a connected electric flux field [1] and they can be utilized with silicon, silicon dioxide, nitride [1]. Materials suitable for piezoelectric are PZT, Lead zirconate titanate and polyvinylidene fluoride (PVDF)

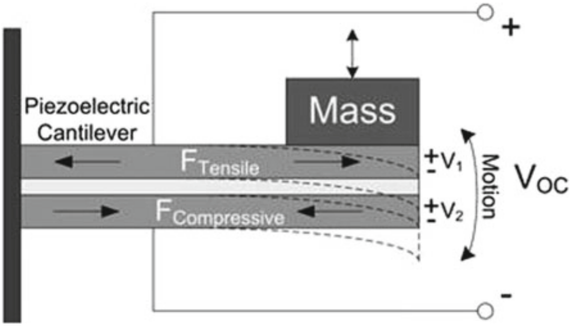


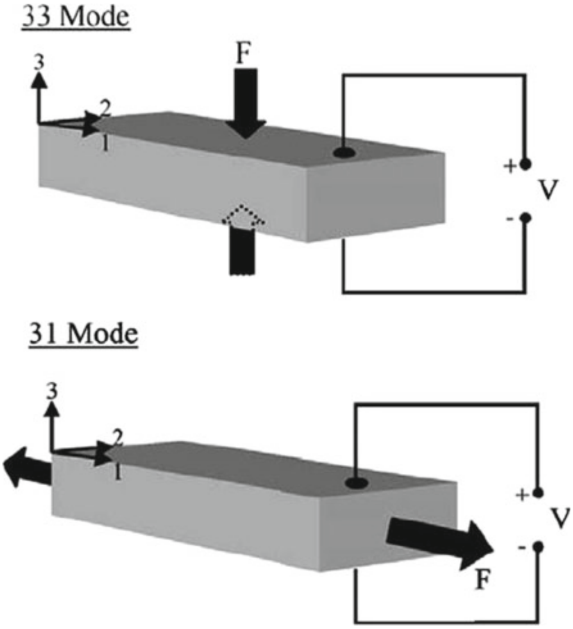
Fig. 1 Piezoelectric- Vintage with cantilever beam causes its crowd through input energy is accelerating. As a result, the beam is deflected, and the material of the beam is loaded with a tensile or compressive force, as a result of which tension is created [10]

polymer [7]. PZT is the best decision for low frequencies. PZT turned out to be the best for optional vibration isolation [11]. Anode introduction adds to the generation of energy, and also the strategy for utilization of the load, tuning and coupling of the resonant frequency [11].

The connection of the material with the electrode determines the system performance of which is parallel (d_{31}) or perpendicular (d_{33}) force, the direction of the input electromagnetic field load [9]. The d_{33} leads to a higher coupling coefficient in comparison with d_{31} communication mode [11] and the possibility of integration with the production of MEM [9]. It also produces a voltage that is higher than the one produced by d_{31} (Fig. 2).

MEMS—Piezoelectric based materials has a substrate of the silicon, layers of constituent electrodes and resistant silicon mass acting on the cantilever at the open end [10]. The main silicon directly receives vibrations from the system environment, so that the intermediate layers receive a transformed force, whereby the cantilever's end parts move in response to the fundamental vibration [10]. By varying the degree of the cantilever, the frequency of the power generator can be adjusted, it can be a disadvantage if a low bandwidth is present [10]. The console serves as a spring-cushion damping system [13]. The bonding pad connects to the load and the voltage signal is measured using an oscilloscope [13]. For an effective generator of resonant frequency energy, the design of the cantilever is of great importance for ensuring consistency between the oscillation frequency and the resonant frequency of the generator [5]. This helps to define the beam's shape and the force applied which depends on the mass produced and the vibration attached to the front end of the rod [5].

Fig. 2 Clutch mode; Description piezoelectric-coefficients relative to the applied deformation direction and the electric field [9]



2 Parameters for the Design

The design for Micropower generator consists of a single Silicon-based frame and cantilevers including various layers including PZT-layers which transform vibrating environment into electricity. The structure is made of a PZT layer between two electrodes and a silicon base layer. PMPG was created from 1.02 μW at a frequency of 20–40 kHz [9]. Since the load is connected to the GM electrode using a vibration sensor, the voltage is produced, this is shown in Fig. 3. The expression for the charge production is as displayed in Eq. 1 [9]. The resonant frequency defers depending on the cantilever deflection. It is however important that cantilevers resonate in frequency same as the surrounding vibration; it is better at a lower frequency.

The deflection in the beam is represented below:

$$m\ddot{z}_0 + (b_e + b_m)\ddot{z}_0 + k\ddot{z}_0 = -m\ddot{z}_1 \tag{1}$$

z_o	Output shift
z_i	Input offset from silicon base
m	mass of bulk silicon
b_e	electrically induced damping factor
b_m	mechanical damping ratio
k	spring constant.

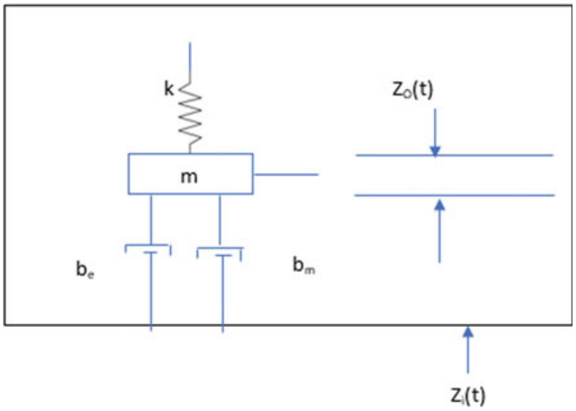


Fig. 3 The incoming energy of the environment causes an oscillation that deflects the beam with a layer of PZT, which leads to a displacement that produces charges [9]

Converted Power becomes [9]:

$$P = \frac{1}{2} b_e \dot{z}_0^2 \tag{2}$$

Deviations from the ambient frequency cause a reduction in output power [9].

$$\text{Resonant frequency,} \quad \omega_n = \sqrt{k/m} \tag{3}$$

From Eq. 3, the body structure by weight of the resonant frequency of the generator. Electrical output power is maximally generated at a resonant frequency is equal to the frequency of the ambient vibration, also in terms of acceleration and input vibration.

$$\begin{aligned} P_{\max} &= P(\omega_n) = \frac{mY^2}{4\zeta} \\ P_{\max} &= \frac{mA^2}{4\xi\omega_n} \end{aligned} \tag{4}$$

where [4, 9]:

A magnitude of the acceleration from the entrance
Y amplitude offset from the input.

As the ambient frequency increases, the displacement amplitude and output power decrease [9].

3 Fabrication Techniques and Processes

The manufacture of the integrated MEMS involves oxidation, doping and etching, which is achieved by forming and shaping silicon or any other material of interest into the desired phase using selected chemical processes and solvents. They are produced as a bulk or as microstructures on silicon, electroplating, lithographic or molding compositions using Atomic Force Microscope (AFM) or Scanning Electron Microscopy (SEM) [3]. There are several manufacturing processes in MEMS whereby it is the desired product that determines the process choice. For the manufacture of cantilever supports commonly used in MEMS equipment, surface micromachining is an effective option. Here, miniature parts are implanted on a silicon substrate using a sacrificial layer, which is etched to obtain the desired shape [3]. The etching process to remove unwanted layer such that it would not affect the patterned device.

The method described above is a subtractive process in which an undesirable part of the structure is removed to give the device the desired size and shape [14] instead of the additive construction process. Piezoelectric generator consists of

5 layers, including PZT and the added mass of silicon [4] which are electrically isolated, but also connected to the upper and lower layers, to serve as electrodes. The electrodes made of aluminium and Ti or Pt are connected to the bond pad. At the open end of the carrier, a large mass of silicon is integrated with a rod made of MEM [4]. From Jing-Quan Liu et al. reports, piezoelectric—generator of 9–25 mm generates about 375 mW of vibration 2.5 m s^{-1} for one cycle of the frequency of 13.7 kHz magnitude [1]. The production process for piezoelectric can be split into 3, embellish that piezoelectric material, micro-console formation [5] and metal mass decomposition. Of great importance also is the material selection in the manufacturing process. The stepwise manufacturing process is discussed below as carried out by Yu, Hua-Bin, Jeon, and others [4, 7, 9].

1. *Sacrificial layer:* the material used may be thermal oxide or chemical vapour deposition. Insulator oxide is used to prevent static friction, the undesirable coincidence of surfaces [3] during tripping, which can lead to residual stresses. This can be accomplished using a silicon wafer on an insulator.
2. *Deposit the electrode:* the lower electrode, Ti or Pt, and the upper electrode on the sol-gel spin coating [4, 9] on the PZT layer using the d_{33} mode PZT. The d_{33} gives a higher voltage than the d_{31} of the same size [9].
3. *Pattern and Etching*: Deep reactive ion etching the structural shape out.
4. *Oxide Deposition*: Plasma Enhancement Chemical vapour deposition is used to precipitate an insulating oxide on which the contacts are located, and the upper part has a pattern and an etching to connect the metal for electrical connection.
5. *Release of the sacrificial layer*: etching is performed to remove the underlying silicon to release the beam. Depending on the sacrificial oxide layer, either reactive ion etching (RIE) or wet etching [3] can be utilised to expose the layer. Using RIE allows immediate release.
6. *Laying test mass*: A mass is applied to the beam, which can be performed using the LIGA method for the poor, SU-8. SU-8 process using photoresist to form a metal structure [3] (Fig. 4).

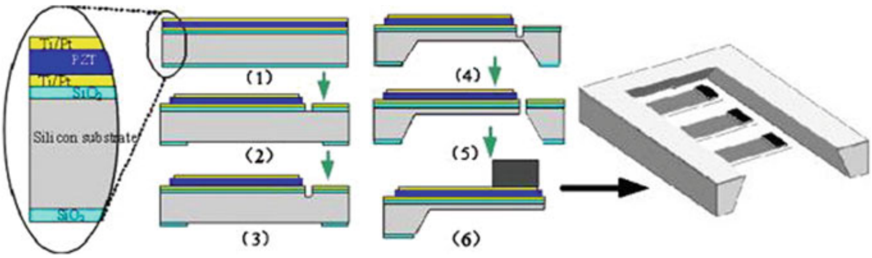


Fig. 4 (1) Applying a beam of PZT-layers on an oxidized silicon substrate (2) and (3) Structuring the beam (4) and (5) Reverse etching of silicon to create a beam of cantilever release (6) Mass deposition [15]

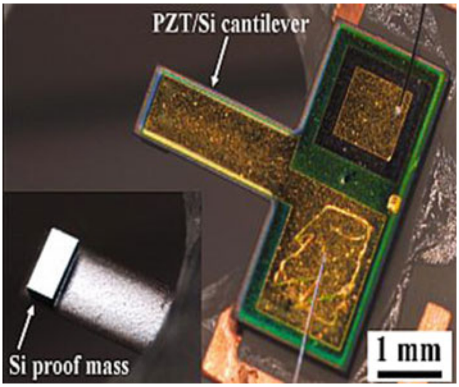


Fig. 5 Deposition of PZT on a console bean for generating electricity [16]

The beam is integrated into a membrane layer consisting of a silicon compound, silicon dioxide or nitride. with a high modulus of elasticity [9]. This is deposited in a single layer (1 0 0) on a silicon wafer using PECVD and thermal oxidation at 750 °C [9]. With solitary [9]. The applied tensile stress is chosen so that after the release a flat cantilever is obtained [9]. The use of thermal silica can cause cantilever bending but with PECVD the curvature decreases [9]. The thin-film piezoelectric Layer can be gotten by RIE using BCI_3:CI_2 [9]. A diffusion barrier to forestall electricity leakage by ZrO_{22} in previous on the lower electrode using a sol-gel spinning which at 350 °C for 1 min and annealed at 700 °C for 15 min Ute s dried stored will [9] (Fig. 5).

4 Technological Benefits of MEMs Piezoelectric

The process made possible energy to be initiated on a larger scale from the environ using solar, thermal and electromagnetic generators. The manufacture of MEMs made it possible to generate energy from the environment in miniature sizes [9]. It is also useful with low power consumption in microwatts. This eliminates the disadvantages of large batteries used for power, and it has a limited lifespan. Packed with ceramic material. Of the three vibration-based energy generation methods that are electrostatic and electromagnetic, a piezoelectric-based generator is more acceptable. Electrostatic generators require input voltage and high output impedance, and high voltage generates a small electrical current [12]. Also, electromagnetic generators have poor planar characteristics coupled with difficulty in integration with MEMS. Thus, piezoelectric material remains an opportunity to extract energy from the environment, since it is compatible with MEMS manufacturing processes [12]. Solar energy was a good factor for the electrical production of electronic devices from the environment, but because of their intermittent supply, it must be stored in a battery.

The piezoelectric cantilever is widely used because it is easy to implement and produce, and because of its high current density, no external voltage is required compared to electrostatic and electromagnetic targets [4. 16]. PZT has a high coefficient of electromechanical adhesion with the effect of the material piezoelectric. Due to the orientation of the axis, it has low dielectric constants, which leads to an increase in the efficiency of energy production [6]. Since no output voltage is required, it can be used anywhere. The contact generates high current with low resistance, which leads to high impedance, which makes it suitable for electronic devices [6]. R and D made it customizable, allowing it to be generated at lower frequencies at 60–200 Hz with higher output power.

Output power refers to the mass volume of silicon, but not with the disadvantage of limited bandwidth, which in most circumstances limits the frequency range, When MEM fabrication is put to use, the generated power is applicable only in miniature integrated devices and for applications with lower power, or for resonance at lower frequencies, a large mass is used. The resonant frequency can be regulated by the offset voltage, which allows damping in the system [15].

5 Feasible Future Application

The frequency of the oscillating environment does not match the resonant frequency of the harvester, which leads to low power. Polymer-based piezoelectric can absorb kinetic energy and exert it on the piezoelectric material, which has a wide frequency range [16]. Using piezoelectric generators in high-frequency environments requires good packaging [8]. In the case of a multi-cantilever, the bandwidth can be increased by adjusting the generator frequency to the resonant frequency of the environment [13]. The flexibility that is typical of harvesters that are polymer-based can be combined with bulk silicon for better performance at low frequency [16]. Converting a low-frequency environment to a higher frequency used by the combine leads to higher productivity [6]. A piezoelectric generator can be combined for best performance with an amplifier and rectifier circuitry. Figure 6 represents a typical micro generator available in the electronic market.

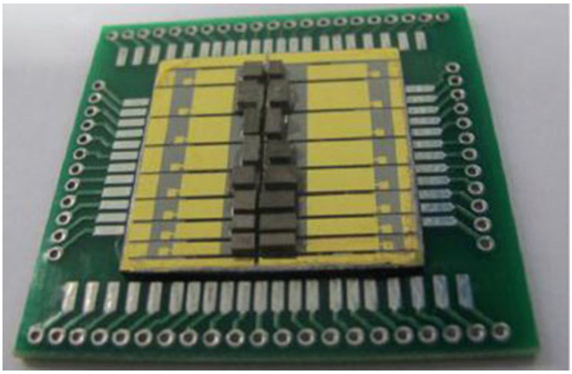


Fig. 6 Microgenerators on board with MEMS piezoelectric

MEMS piezoelectric explores a structure in which energy is continuously integrated into electronics to achieve ultra-low integration power. The manufacturing process of MEMS made it possible and in embedded systems and was accompanied by its power conditioning circuit which can be used to generate the energy maximally [4].

6 Conclusion

Technological advancement had recently been making advances at using MEMs to manufacture slimmer, smaller, sleeker, tinier, lighter, more power sufficient, lower energy consuming and more economical (cheaper) devices. To solve the difficulty of changing batteries for wireless electronics, sensors especially in the remote areas, there is need for self-energy producing devices to give constant supply of power. Among other options of MEMs based generator are electrostatic generation, dielectric elastomers and piezoelectric materials all due to their abilities to convert an environmental phenomenon into electrical energy and also ability to be integrated into electronic devices. Generally, MEM based generators are useful for providing sources of power in small wireless sensor nodes in place of conventional polar batteries. Wireless sensors for example, can be powered using energy harvested from vibrations environment. Also, other MEM harvestable sources can be through wind, power flow, electromagnetic waves, and solar photons all from the environment of the functioning device.

References

1. Basin-Gharb N (2008) Piezoelectric MEMS: materials and devices. In: Piezoelectric and acoustic materials for transducer applications. ©Springer, pp 413–430
2. Gad-el-Hak M (2006) MEMS applications. CRC Press, pp 1–3 (Chapter 1)
3. Eisele H (2013) ELEC5500 MEMS and NEMS lecture notes, 2013–2014
4. Yu H, Zhou J, Deng L, Wen Z (2014) A vibration-based MEMS piezoelectric energy harvester and power conditioning circuit. In: NCBI, 19 Feb 2014, pp 3323–3341
5. Saadon S, Sidek O (2013) Shape optimization of cantilever-based MEMS piezoelectric energy harvester for low frequency applications. In: 2013 UKSim 15th international conference on computer modelling and simulation (UKSim). ©IEEE, Cambridge, 10–12 Apr 2013, pp. 202–208
6. Kim M, Hwang B, Ham YH, Jeong J, Min NK, Kwon KH (2012) Design, fabrication, and experimental demonstration of a piezoelectric cantilever for a low resonant frequency microelectromechanical system vibration energy harvester. J Micro/Nanolith MEMS MOEMS 11(3):033009
7. Hua-Bin F, Jing-Quan L, Zheng-Yi X, Lu D, Di C, Bing-Chu C, Yue L (2006) A MEMS-based piezoelectric power generator for low frequency vibration energy harvesting. Chin Phys Lett 23(3) © IOP Publishing
8. Amirtharajah R, Chandrakasan AP (1998) Self-powered signal processing using vibration-based power generation. IEEE J Solid-State Circ 33(5):687–695

9. Roundy S, Wright PK, Rabaey J (2003) A study of low-level vibrations as a power source for wireless sensor nodes. Computer Communications 26(11):1131–1144
10. Torres EO, Rincon-Mora GA (2014) Energy-harvesting chips and the quest for everlasting life. In: Georgia tech analog and power IC design lab. © EE Times [Online], 27 May 2014. Available: http://www.eetimes.com/document.asp?doc_id=1273025
11. Auton SR, Sodano HA (2007) A review of power harvesting using piezoelectric materials (2003–2006). Smart Mater Struct 16(3)
12. Cook-Chennault KA, Thambi N, Sastry AM (2008) Powering MEMS portable devices—a review of non-regenerative and regenerative power supply systems with special emphasis on piezoelectric energy harvesting systems. Smart Mater Struct 17. ©IOP publishing
13. Ramadan KS, Sameoto D, Evoy S (2014) A review of piezoelectric polymers as functional materials for electromechanical transducers. Smart Mater Struct 23(3). ©IOP Publishing
14. Brydson RM, Hammond C (2005) Generic methodologies for nanotechnology: classification and fabrication. In: Nanoscale science and technology, Wiley
15. Liu, J-Q, Fang H-B, Xu Z-Y, Mao X-H, Shen X-C, Chen D, Liao H, Cai B-C (2008) A MEMS-based piezoelectric power generator array for vibration energy harvesting. Microelectron J 39(5):802–806. ©Elsevier
16. EPFL [Online] Piezoelectric MEMS Vibration Energy Harvesters. Available: https://www.epfl.ch/labs/lmts/lmts-research/enviromems/page-129700-en-html/

Face Recognition Using Laplacian Completed Local Ternary Pattern (LapCLTP)

Sam Yin Yee, Taha H. Rassem, Mohammed Falah Mohammed and Suryanti Awang

Abstract Nowadays, the face is one of the typical biometrics that has high-security technology in the biometrics field. In face recognition systems, feature extraction is considered as one of the important steps. In feature extraction, the important and interesting parts of the image are represented as a compact feature vector. Many features had been proposed in the image processing fields such as texture, colour, and shape. Recently, texture descriptors are playing an important and significant role as a local descriptor. Different types of texture descriptors had been proposed and used for face recognition task, such as Local Binary Pattern (LBP), Local Ternary Pattern (LTP), and Completed Local Ternary Pattern (CLTP). All these texture features have achieved good performances in terms of recognition accuracy. In this paper, we propose to improve the performance of the CLTP and use it for face recognition. A Laplacian Completed Local Ternary Pattern (LapCLTP) is proposed in this paper. The image is enhanced using a Laplacian filter for pre-processing image process before extracting the CLTP. JAFFR and YALE standard face datasets are used to investigate the performance of the LapCLTP. The experiment results showed that the LapCLTP outperformed the original CLTP in both datasets and achieved higher recognition accuracy. The LapCLTP achieved 99.24%, while CLTP achieved 98.78% with JAFFE dataset. IN YALE, the LapCLTP achieved 85.13%, while CLTP, only 84.46%.

Keywords Face recognition · Completed local ternary pattern (CLTP) · Laplacian filter · (LAPCLTP) · Image classification · Face database

S. Y. Yee · T. H. Rassem (✉) · M. F. Mohammed · S. Awang
Faculty of Computing, College of Computing and Applied Sciences,
Universiti Malaysia Pahang, 26300 Kuantan, Malaysia
e-mail: tahahussein@ump.edu.my

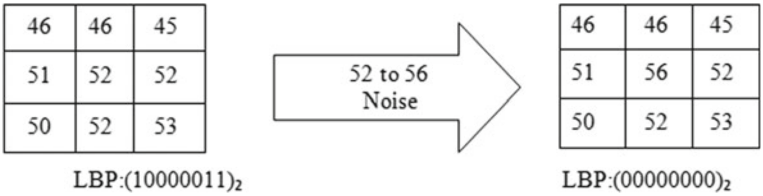
Z. Zakaria and R. Ahmad (eds.), *Advances in Electronics Engineering*,
Lecture Notes in Electrical Engineering 619,
https://doi.org/10.1007/978-981-15-1289-6_29

1 Introduction

Nowadays, a face recognition system is very useful because it can be used as authentication and security. Many systems are using different ways to identify a person such as an iris pattern detection [1], voice recognition [2], face recognition [3], and fingerprint system [4]. However, the friendliest and normal system is the identification of a person using a face. The face recognition system should be able to recognize the faces even with different facial expressions, wearing a glass, changing hairdo, etc. The process of face recognition system begins by distinguishing the presence of a face in an image. Mostly, a face detection process can decide whether an image contains a face or not. Otherwise, the role of the system is to find a position of at least one face in the image. After detecting a face in the image, the next step is to extract the facial features. Many of features are studied and investigated in the literature review.

Texture descriptors have shown good performances as powerful features that can be used to distinguish and classify different image types. Different types of the texture descriptors have been proposed during the last decade. These types are used for different tasks in the image processing and computer vision fields. Local Binary Pattern (LBP) is one of the famous texture descriptors that was proposed for texture image classification first, and then used for different types of image classifications [5, 6]. After that, different types of texture descriptors are inspired from the LBP such as Local Ternary Pattern (LTP) [7], Completed Local Binary Pattern (CLBP) [8], and Completed Local Ternary Pattern (CLTP) [9]. The LBP was reported to be initially used for texture descriptors in 2002 [5]. The superiorities of LBP are bound to its invariance to revolution, robustness against monotonic dim level change, and its low computational complexity, but it also has some disadvantages, which are high-sensitivity to noise as well as sometimes not being able to distinguish different patterns [10]. An example of LBP sensitivity to noise can be shown in Fig. 1, while an example of the second disadvantage can be shown in Fig. 2. In Fig. 2, different patterns of LBP might be encoded into a similar LBP code that decreases the LBP distinguish property.

Despite that, the Local Ternary Pattern (LTP) has been proposed to overcome the drawback of LBP which is more robust to noise compared with LBP. However, the LTP is still suffering from the second problem. There are many texture descriptors that have been proposed afterwards, such as CLBP, CLBC, and



46	46	45
51	52	52
50	52	53

LBP:$(10000011)_2$

46	46	45
51	56	52
50	52	53

LBP:$(00000000)_2$

Fig. 1 LBP sensitive to noise drawback

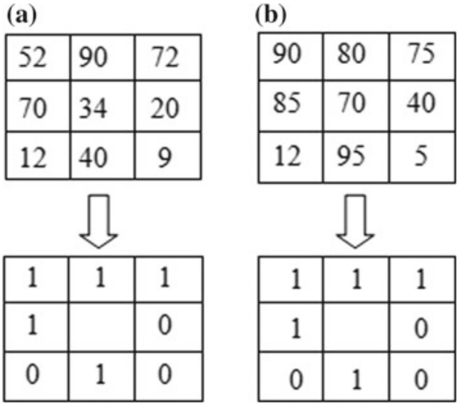


Fig. 2 A different Pattern encoded to the same LBP code drawback

CLTP. The Completed Local Ternary Pattern (CLTP) [9] showed more robustness to noise than the existing texture descriptor. Moreover, it outperformed the previous texture descriptors in many fields [10–12].

In this paper, a Laplacian Completed Local Ternary Pattern (LapCLTP) is proposed. In LapCLTP, the CLTP is extracted after applying the Laplacian filter into the image. Laplacian filter is good to use for searching the fine details of an image which can bring more details features that lead to higher accuracy result.

The present paper is organized as follows. Section 2 discusses the related work where some of existing texture descriptors are explained. Section 3 presents the proposed LapCLTP as well as the general structure of the proposed face recognition system. Then Sect. 4 presents the experiment and results of the LapCLTP for face recognition task. Finally, Sect. 5 concludes the paper.

2 Related Work

Texture descriptors are playing an important role in many image processing fields. LBP and its variants are mostly used in many applications. To extract these texture descriptors, the size of the texture pattern is important and can be affect the final results. A texture pattern radius size is allocated to carry out the extraction process in the earlier stages of feature extraction.

The widely used texture pattern sizes are (1, 8), (2, 16), and (3, 24). These radius sizes are shown in Fig. 3.

First, a set of training and testing images are prepared and processed for feature detection. Feature detection is to determine whether or not an image has a face. Once the target places of the face are detected, it will proceed to face extraction step to extract the facial characteristics. The feature extraction stage is the process to extract the target features from each image i.e., training and testing images. After proceeding with the feature extraction, the similar training features of images are

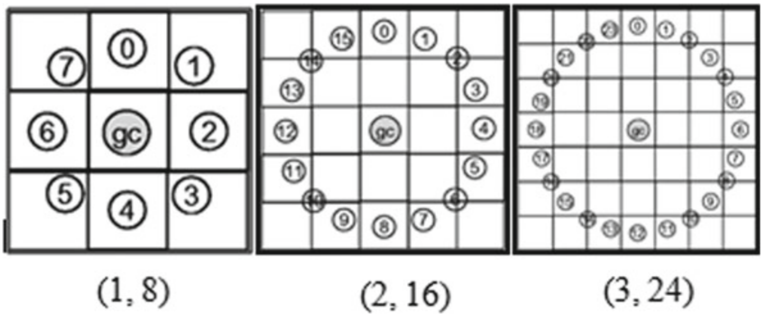


Fig. 3 The illustration of different pattern texture radius sizes [13]

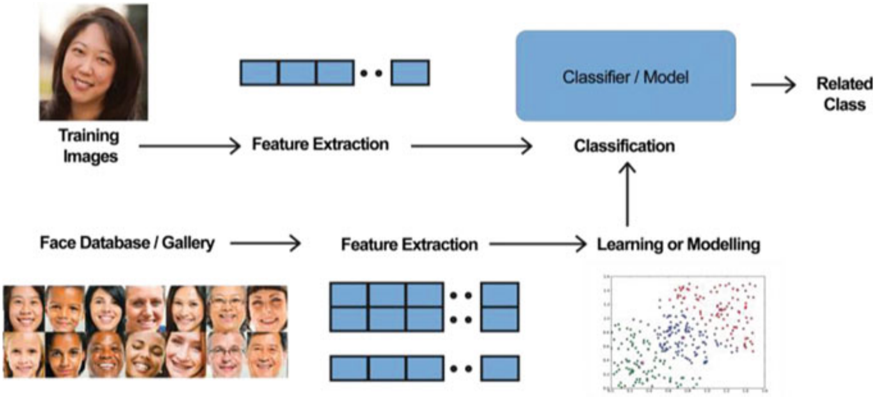


Fig. 4 General face recognition structure

grouped together. Then, the features from the testing images will be extracted and compared with the extracted features of the training image to find the smallest distance or closest value between the training and testing image features. The general face recognition structure can be shown in Fig. 4.

Many of texture descriptors are utilized to perceive and distinguish face, such as Local Binary Pattern (LBP), Local Ternary Pattern (LTP), and Completed Local Ternary Pattern (CLTP).

2.1 Local Binary Pattern (LBP)

The Local Binary Pattern (LBP) was presented by Ojala et al. in 1996 [5]. LBP is an effective device to portray the neighbourhood grey-level attributes of surface composition. LBP surface operator has become a popular methodology in different applications. It can be classified as a method to deal with the generally different factual and auxiliary models of texture investigation.

The LBP has the capability to extract the uniform pattern by extracting the data information of the image to be more precise in each image pixel. As a result, LBP is efficient for arranging the neighbourhood spatial structure of a picture. The LBP feature descriptors are improvised to utilize neighbourhoods of various sizes [6] by using a roundabout neighbourhood introducing values at non-whole number pixel that organizes permits to any sweep and pixels in the area. Moreover, LBP is a very simple texture feature that is used for image processing by converting the image pixel and using the thresholding to the neighbourhood of each pixel and comparing it with the value of the central pixel and binary number followed by decimal number as the outcome. LBP computation is shown in Eq. 1.

$$LBP_{P,R} = \sum_{p=0}^{P-1} 2^P s\left(i_p - i_c\right), \quad s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{1}$$

LBP is suffering from two drawbacks as shown in Figs. 1 and 2. Due to those drawbacks, many different texture descriptors are proposed and inspired from the LBP after overcoming its drawbacks.

2.2 Local Ternary Pattern (LTP)

In a way to enhance the robustness of neighbourhood code, the Local Binary Pattern (LBP) texture descriptors are supplanted by a Local Ternary Pattern (LTP) texture features. In LTP, the user can set the number of the threshold [7]. This might make the LTP code to be more impervious to noise. However, the changes in grey-level will no more be entirely invariant.

The pixel contrasts between the neighbouring pixels and the inside pixel have been encoded by LTP. This process can be done by comparing the difference of the middle pixel and P neighbours on a circle of span R. The LTP can be calculated as shown in Eq. 2.

$$LTP_{P,R} = \sum_{p=0}^{P-1} 2^P s\left(i_p - i_c\right), \quad s(x) = \begin{cases} 1, & t \geq 0 \\ 0, & -t < x < t \\ -1, & x < -t \end{cases} \tag{2}$$

2.3 Completed Local Ternary Pattern (CLTP)

CLTP features descriptor has been proposed by Rassem and Khoo [9] for rotation invariant texture classification. There are three components of CLTP, which are CLTP_S, CLTP_M, and CLTP_C.

Firstly, two signs, namely an upper sign and lower sign are computed as shown in Eqs. 3 and 4.

$$s_p^{upper} = s\big(i_p - (i_c + t)\big), \quad s_p^{upper} = s\big(i_p - (i_c - t)\big) \tag{3}$$

$$m_p^{upper} = \big|i_p - (i_c + t)\big|, \quad m_p^{lower} = \big|i_p - (i_c - t)\big| \tag{4}$$

To get the outcome of CLTP_S, the upper sign and lower sign components need to be calculated by converting the pixel of the image to a binary code. After that, the binary codes of CLTP_S$_{\mathrm{P,R}}^{\mathrm{upper}}$ and CLTP_S$_{\mathrm{P,R}}^{\mathrm{lower}}$ are combined or added together to get the CLTP_S feature. The CLTP_S can be calculated as shown below.

$$CLTP\_S_{P,R}^{upper} = \sum_{p=0}^{P-1} 2^P s\big(i_p - (i_c + t)\big), \quad S_P^{upper} = \begin{cases} 1, & i_p \geq i_c + t, \\ 0, & otherwise, \end{cases} \tag{5}$$

$$CLTP\_S_{P,R}^{lower} = \sum_{p=0}^{P-1} 2^P S\big(i_p - (i_c - t)\big), \quad S_P^{lower} = \begin{cases} 1, & i_p < i_c - t \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

The two operators, then, are concatenated to form CLTP_S$_{\mathrm{P,R}}$ as follows:

$$CLTP\_S_{P,R} = \left[CLTP\_S_{P,R}^{upper} \quad CLTP\_S_{P,R}^{lower} \right] \tag{7}$$

Here the illustration of CLTP_S process is shown in Fig. 5.

CLTP_M is similar to the CLTP_S, which needs to add the CLTP_M$_{\mathrm{P,R}}^{\mathrm{upper}}$ and CLTP_M$_{\mathrm{P,R}}^{\mathrm{lower}}$ to get the CLTP_M$_{\mathrm{P,R}}$. The output of CLTP_M will also be in a binary form same as CLTP_S as shown below:

$$\begin{aligned} CLTP\_M_{P,R}^{upper} &= \sum_{p=0}^{P-1} 2^P t\left(m_p^{upper}, c\right), \quad t\left(m_p^{upper}, c\right) \\ &= \begin{cases} 1, & \big|i_p - (i_c + t)\big| \geq c, \\ 0, & \big|i_p - (i_c + t)\big| < c, \end{cases} \end{aligned} \tag{8}$$

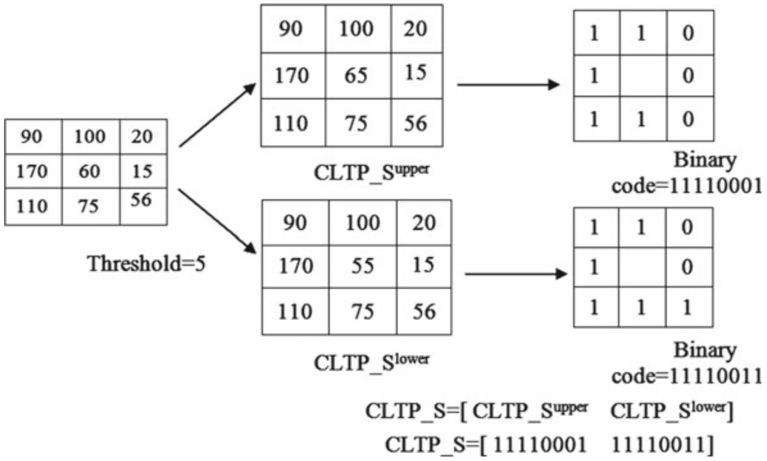


Fig. 5 CLTP_S operator

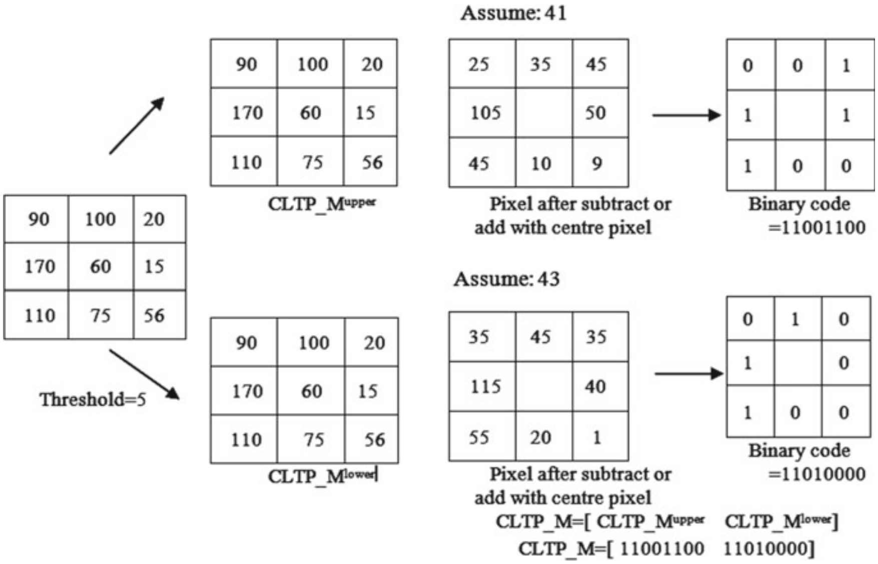


Fig. 6 CLTP_M operator

$$\begin{aligned} CLTP\_M_{P,R}^{lower} &= \sum_{p=0}^{P-1} 2^P t(m_p^{upper}, c), \quad t\left(m_p^{lower}, c\right) \\ &= \begin{cases} 1, & \left|i_p - (i_c - t)\right| \geq c, \\ 0, & \left|i_p - (i_c - t)\right| < c, \end{cases} \end{aligned} \tag{9}$$

$$CLTP\_M_{P,R} = \left[CLTP\_M_{P,R}^{upper} \quad CLTP\_M_{P,R}^{lower}\right] \tag{10}$$

The illustration of CLTP_M process in Fig. 6.

The outcome of CLTP_C is a 2D matrix of the binary value. It compares the average pixel number of the original pixel numbers with the upper and lower signs of CLTP_C component. Then, it comes out with a 2D matrix of binary value of $\mathrm{CLTP\_C}_{\mathrm{P,R}}^{\mathrm{upper}}$ and 2D matrix of a binary value of $\mathrm{CLTP\_C}_{\mathrm{P,R}}^{\mathrm{lower}}$. The mathematical method is shown as Eqs. 11 and 12:

$$CLTP\_C_{P,R}^{upper} = t\left(i_c^{upper}, c_I\right) \tag{11}$$

$$CLTP\_C_{P,R}^{lower} = t\left(i_c^{lower}, c_I\right) \tag{12}$$

3 Proposed Laplacian Completed Local Ternary Pattern (LapCLTP)

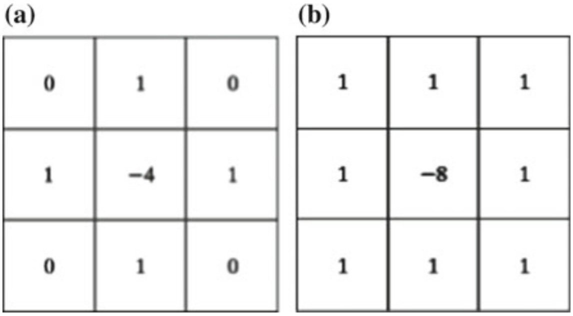
3.1 Laplacian Filter

The Laplacian operator is a case of second subordinate strategy for an upgrade. It is used to sharpen the images and it is especially great at disclosing the fine detail in a picture. Any element with noise can be enhanced by using a Laplacian filter. Thus, one application of a Laplacian operator is to re-establish fine detail to a picture, which has been smoothed to remove noise.

The Laplacian L(x, y) of a picture with pixel intensity values I(x, y) is given as follow in Eq. 13.

$$L(x, y) = \frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2} \tag{13}$$

It is very useful to feature dark level discontinuities in a picture and attempt to deemphasize locales with gradually differing dim levels. This activity will create the pictures, which have grayish edge lines and different discontinuities on a featureless background. Laplacian operator has a positive and negative operator. It uses it to produces the internal and outer edges of images, respectively. Besides that, there are two filters that usually utilized the 3 × 3 kernel as shown in Fig. 7.



(a)

0	1	0
1	−4	1
0	1	0

(b)

1	1	1
1	−8	1
1	1	1

Fig. 7 Laplacian usually used **a** filter invariant to 90° rotations. **b** Filter invariant to 45° rotations

3.2 Proposed LapCLTP

The LapCLTP will be extracted from all training and testing images.

Figure 8 summarizes the LapCLTP extraction process.

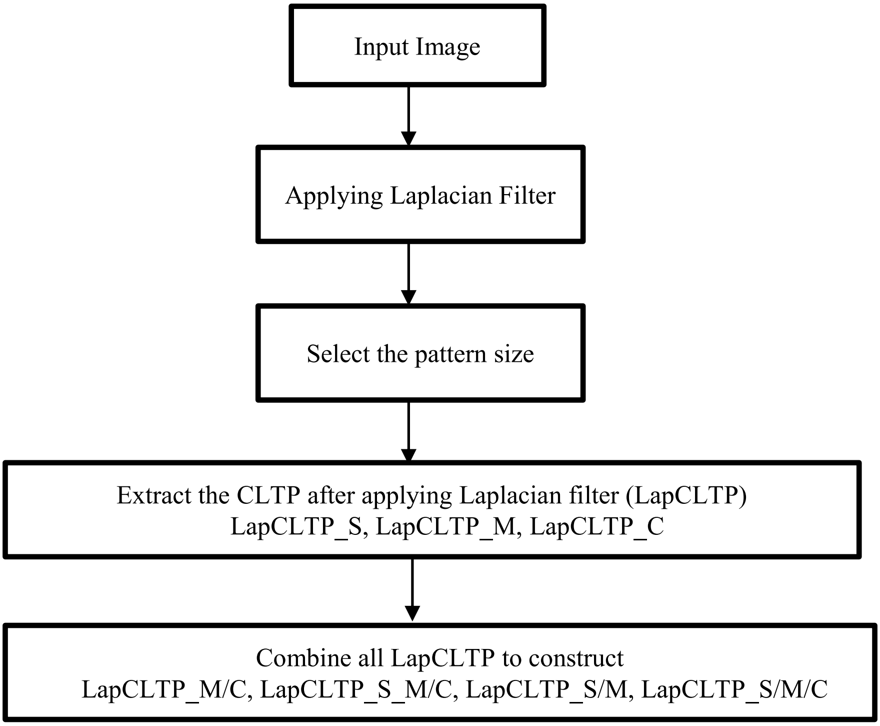

Input Image
Applying Laplacian Filter
Select the pattern size
Extract the CLTP after applying Laplacian filter (LapCLTP)
LapCLTP_S, LapCLTP_M, LapCLTP_C
Combine all LapCLTP to construct
LapCLTP_M/C, LapCLTP_S_M/C, LapCLTP_S/M, LapCLTP_S/M/C

Fig. 8 The proposed LapCLTP extraction process

4 Experiments and Results

To evaluate the LapCLTP performance for face recognition task, JAFFE and YALE standard image datasets are used.

4.1 *Japanese Female Facial Expressions Database (JAFFE)*

JAFFE database [14] comprised of 213 face images from 10 different classes of Japanese females in Japan. The features extraction of (1, 8), (2, 16), and (3, 24) radius sizes, based on 2, 5, and 10 random training images from the classes, were used. Each class has 20 JPEG images with a different view of facial expressions including angry, smile, sad, worry, nervous, neutral, and others. The JAFFE image size is 256 × 256.

Examples of JAFFE images are shown in Fig. 9.

Table 1 shows the face recognition accuracy results of CLTP and the proposed LapCLTP. The table shows the experiment results for different radiuses (1, 8), (2, 16), (3, 24) and different training images (= 2, 5, 10). All the combinations of CLTP and LapCLTP are evaluated under the above condition. The CLTP showed good performance and outperformed the LapCLTP in some cases as shown in the table. However, the performance of the proposed LapCLTP was increased and outperformed the original CLTP at the end. The LapCLTP_S/M/C$_{3,24}$ achieved recognition rate using JAFFE, and reached 99.24% compared with 98.78% by CLTP_S/M/C$_{3,24}$ when 10 training images were used. Generally, the LapCLTP performed better than CLTP in many cases using JAFFE dataset with various training images of N.

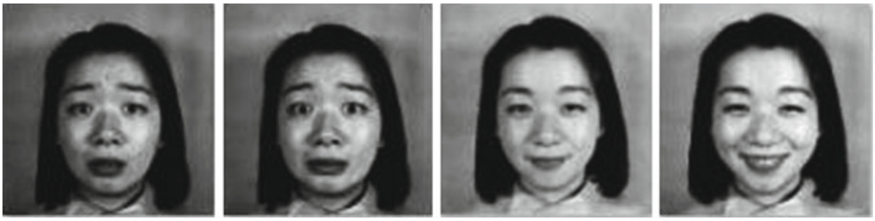


Fig. 9 Some images from JAFFE database

Table 1 Classification accuracy (%) on JAFFE database (JAFFE)

JAFFE database	R = 1, P = 8			R = 2, P = 16			R = 3, P = 24		
	2	5	10	2	5	10	2	5	10
CLTP_S	71.11	**79.00**	**83.62**	**78.33**	**84.37**	87.68	**81.06**	**88.43**	**91.94**
LapCLTP_S	**71.72**	77.35	81.94	77.17	84.07	**88.09**	79.87	85.58	89.50
CLTP_M	74.07	82.12	86.59	78.25	83.24	85.86	75.60	82.47	86.32
LapCLTP_M	**75.26**	**83.82**	**88.67**	**79.93**	**85.93**	**89.84**	**81.89**	**87.95**	**91.15**
CLTP_M/C	82.88	90.17	93.64	88.17	93.77	96.27	89.49	94.80	96.77
LapCLTP_M/C	**88.51**	**93.77**	**96.66**	**91.31**	**95.48**	**96.80**	**91.69**	**96.13**	**98.40**
CLTP_S_M/C	84.71	91.55	95.13	90.55	94.89	96.93	90.44	94.75	96.95
LapCLTP_S_M/C	**89.04**	**94.27**	**96.52**	**91.61**	**96.17**	**97.98**	**91.87**	**96.31**	**97.93**
CLTP_S/M	80.43	**88.33**	**92.07**	85.94	91.69	**95.02**	86.72	92.05	**95.89**
LapCLTP_S/M	**82.23**	87.11	90.27	**86.00**	**91.87**	94.52	**87.09**	**92.10**	95.42
CLTP_S/M/C	87.09	92.97	96.08	**92.09**	95.78	97.82	**93.26**	97.14	98.78
LapCLTP_S/M/C	**91.27**	**95.64**	**97.40**	92.04	**96.17**	**98.32**	92.21	**97.15**	**99.24**

4.2 *Yale Face Database (YALE)*

YALE database [15] was captured from 15 people, everyone is requested to capture 11 images, so, it contains 165 images in the whole database. Each image is captured from alternate points of view, with a big difference in brightening and face expression. Each image in the YALE database was physically trimmed by following the face detected point sizes. Some images from the database are shown in Fig. 10

The classification results of YALE Face Database for = (2, 5, 10) are shown in Table 2. The LapCLTP_S/M/C$_{2,16}$ has achieved the best result with 85.53% using 10 training images compared with 84.46% as the best CLTP result achieved by CLTP_S/M/C$_{3,24}$. The Performance of CLTP and LapCLTP were interchangeable. Generally, the LapCLTP was able to achieve a high recognition rate in YALE database.

Fig. 10 Some images from YALE database

Table 2 Classification Accuracy (%) on YALE Face Database (YALE)

YALE database	R = 1, P = 8			R = 2, P = 16			R = 3, P = 24		
	2	5	10	2	5	10	2	5	10
CLTP_S	45.50	58.11	**70.20**	50.20	63.17	71.47	56.72	67.25	72.40
LapCLTP_S	**48.62**	**58.38**	62.20	**56.72**	**67.25**	**72.40**	**64.01**	**72.75**	**76.86**
CLTP_M	**55.48**	**66.03**	**68.07**	**61.23**	**72.26**	75.33	**65.30**	**77.41**	**80.87**
LapCLTP_M	52.38	60.78	65.93	56.72	68.51	**77.467**	61.47	71.41	69.09
CLTP_M/C	63.47	73.26	74.26	**71.75**	78.84	78.40	**73.11**	79.70	83.47
LapCLTP_M/C	**64.10**	**74.24**	**78.60**	71.28	**81.89**	**85.53**	72.29	**79.96**	**83.87**
CLTP_S_M/C	64.61	75.14	76.73	71.12	**79.57**	80.73	72.46	80.26	80.47
LapCLTP_S_M/C	**66.97**	**76.51**	**78.87**	**72.10**	78.52	**84.67**	**74.07**	**80.82**	**81.00**
CLTP_S/M	**61.30**	**71.22**	**70.93**	**67.40**	**77.36**	79.24	**69.41**	**82.94**	**84.46**
LapCLTP_S/M	60.24	70.78	69.53	66.73	75.95	**80.93**	69.13	77.65	81.73
CLTP_S/M/C	63.76	75.14	73.26	**74.25**	**83.13**	83.20	**76.61**	84.07	82.80
LapCLTP_S/M/C	**70.44**	**79.00**	**81.27**	74.07	78.56	**85.53**	75.25	**84.30**	**85.13**

5 Conclusion

In this paper, the Laplacian Completed Local Ternary Pattern (LapCLTP) is proposed to enhance the performance of CLTP in face recognition. The Laplacian filter is added during the extraction process to get the LapCLTP. The Laplacian filter can find more fine details in the image as well as helping to remove some noises. The proposed LapCLTP is tested for face recognition task using JAFFE and YALE standard face image datasets. The LapCLTP achieved 99.24% and 85.13% as the highest recognition rates in JAFFE and YALE datasets, respectively. The LapCLTP performance is tested under different number of training images and different texture pattern sizes. The texture pattern size can affect the recognition accuracy, where in YALE dataset, the best results were achieved by using radius 2 and neighbours 16 pixels. In JAFFE, the LapCLTP achieved better results with radius 3 and 24 neighbor pixel texture patterns. In the future work, the performance of the LapCLTP will be investigated for face recognition task using more face datasets.

Acknowledgements This research has been supported by Universiti Malaysia Pahang (UMP) and it was funded by UMP Grant Number RDU160104 and RDU180365.

References

1. Bhattacharyya D, Das P, Bandyopadhyay S (2008) IRIS texture analysis and feature extraction for biometric pattern recognition. Sersc.Org, pp 53–60
2. Hecker MH (1971) Speaker recognition. An interpretive survey of the literature. In: ASHA monographs, vol 16, American Speech and Hearing Association, pp 1–103

3. Bledsoe WW (1964) The model method in facial recognition. In: Panoramic Research, California: Panoramic Research, Inc
4. Hankley WJ, Tou JT (1968) Automatic fingerprint interpretation and classification via contextual analysis and topological coding. Ohio State University
5. Ojala T, Pietikäinen M, Harwood D (1996) A comparative study of texture measures with classification based on featured distributions. Pattern Recogn 29(1):51–59
6. Ojala T, Pietikäinen M, Mäenpää T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell 24 (7):971–987
7. Tan X, Triggs B (2010) Enhanced local texture feature sets for face recognition under difficult lighting conditions. IEEE Trans Image Process 19(6):1635–1650
8. Guo Z, Zhang L, Zhang D (2010) A completed modeling of local binary pattern operator for texture classification. IEEE Trans Image Process 19(6):1657–1663
9. Rassem TH, Khoo BE (2014) Completed local ternary pattern for rotation invariant texture classification. Sci World J 2014:10
10. Rassem TH, Makbol NM, Yee SY (2017) Face recognition using completed local ternary pattern (CLTP) texture descriptor. Int J Electr Comput Eng 7(3):1594–1601
11. Rassem TH, Khoo BE, Makbol NM, Alsewari ARA (2017) Multi-scale colour completed local binary patterns for scene and event sport image categorisation. IAENG Int J Comput Sci 44(2):197–211
12. Rassem TH, Mohammed MF, Khoo BE, Makbol NM (2015) Performance evaluation of completed local ternary patterns (CLTP) for medical, scene and event image categorisation. In: 2015 4th international conference on software engineering and computer systems, ICSECS 2015: virtuous software solutions for big data, pp 33–38
13. Doost HE, Amirani M (2013) Texture classification with local binary pattern based on continues wavelet transformation. Int J Adv Res Electr Electron Instrum Eng 2(10):4651–4656
14. Lyons MJ (1999) Automatic classification of single facial images. IEEE Trans Pattern Anal Mach Intell 21(12):1357–1362
15. Belhumeur PN, Hespanha JP, Kriegman DJ (1997) Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Trans Pattern Anal Mach Intell 19(7):711–720

An Insider Threat Factors and Features Categorization for Manufacturing Execution System

Warusia Mohamed Yassin, Rabiah Ahmad and Nur Ameera Natasha Mohammad

Abstract An insider threats turn cyber world into insecure data breaches and system compromised as the insider having legitimate access to information of critical assets. Furthermore, the threat reflected unnoticeable and none able to foresee what, when and how literally the trusted insiders who has authority launched the threats against an organization. Due to this, there is lack of theoretical view discussion by the research community that can be used as a reference to categorize factors, specifically features that can contribute to the insider threats in manufacturing execution systems (MES). Therefore, a theoretical view to categorize factors and features which represent the behavior of insider threats in MES is proposed based on conducted literature survey. These threats could be grouped into three major factors i.e. human, systems and machine as stressed, and consequently a possible feature that can be a contributor for every single factor identified based on previous researcher recommendations. For the purpose of facilitate the understanding, the real scenario from the automation execution system from manufacturing sector is chosen as case study. Each factor and every single related feature identified, grouped and fact been highlighted. Hence, a theoretical framework for MES could be derived and facilitate as a standard guideline to mitigate insider threats in manufacturing field.

Keywords Insider threats · Manufacturing execution systems · Mitigate · Insider threat factors · Insider threat features

W. M. Yassin (✉) · R. Ahmad · N. A. N. Mohammad
Center for Advanced Computing Technology (C-ACT), Fakulti Teknologi Maklumat Dan Komunikasi (FTMK), Universiti Teknikal, Melaka, Malaysia
e-mail: s.m.warusia@utem.edu.my

Z. Zakaria and R. Ahmad (eds.), *Advances in Electronics Engineering*,
Lecture Notes in Electrical Engineering 619,
https://doi.org/10.1007/978-981-15-1289-6_30

1 Introduction

An insider usually a privileged user who possess authorized access to confidential assets or information's, and they well acknowledge on installed systems and business processes weaknesses. Insider threats can be defined as an attacker from part of the organizations whether current or former employee, trusted business partner, supplier, contractor or subcontractor who has or has had authorized access to that organization's IT assets and launch threats that can give harm or negative impact on the information security elements of the organizations [3]. Besides, an insider attacks a.k.a. threats are more complex and harder to be detectable as contrast to external attacks whereas their marks are easy to track [12]. Furthermore, there is increasing in number of insider threats recently and to be handling such threats keep growing and turn to be highly important. The insider threats generally affect the breaches of critical part of information security as reported in and directly causes a serious damage to the organization i.e. loss of revenue, a reduction in productivity, cost arising, bankruptcy and negative impact on business activity. The notorious example of infected field by an insider threats most likely are in manufacturing. Today resolution in manufacturing are more concern in manufacturing execution systems (MES) as to carry out the task of management planning, production order processing as well as production control and monitoring. However, despite everything the security concern are the great challenges and risk particularly, if MES utilizing a cloud platform to execute their operations. Furthermore, organization will suffer the loss of control over important data and admit that only the cloud provider able to deal with any issues including security as the organization totally tied with cloud business processes. Furthermore, manufacturing businesses heading into innovative age called fourth industrial revolution (Industry 4.0) which more related to high-tech of automation process, internet and sensors technology, concurrent digital and intelligent data exchange. This directly involves the use of MES and its facility as a whole and now most companies are beginning to realize the benefit of such system. However, even though MES getting much popularity, but it is also suffering with issues of insider threats. As the insider have a privileged account and legitimate access, the difficulties in identifying it remain as open challenges. Besides, there is no theoretical view which can be referred to categorize the factor which may contribute to an insider threat. An example of impact is more concern on production line effected, production time (downtime) and quality control infected. Therefore, the factors and features that contribute to insider threats need to be identified as a reference by industry especially that may assist in the designing of the appropriate security framework. As such, in this paper, a factors and features for insider threat with existing real scenario from manufacturing, chosen to illustrate how the insider threats factors and its characteristic may identified and classified.

1.1 An Insider Threat Factors

Based on several literature survey, source of insider threat activity can be divided into three main factors namely human, machine and systems perspective. In human factors perspective, most of the threat happened for the self-satisfaction of insider which he/she used privacy information of the organization gathered from their own effort easily as they already have legitimate access to enter the organization. The biggest challenge for each of the organizations in this world has to face is trustworthy issues towards their employers. Other than that, the effectiveness of the organization's management also plays as one of the big part in order to tackle insider threats [8]. The identified features under this factors are disgruntlement, anger management issues [4], socially mean character, self-promotion and emotion sensitivity [9]. Besides, the system factors itself have legitimate data or record which help insider to enter or access into the organization and easily gathered what they want which to launch the threat. In addition, a number of researcher [8, 14, 16, 20] has highlight the features that influence the insider threat under the system factors i.e. server operation, vulnerability of defense devices, false injection attacks, physical fault via smart grid technologies, electrical power systems, disruptions of sensors and actuators. Other than these two factors, machine factors also play a dominant contributor for insider threats. Therefore, choosing the right machine with latest security features can help to mitigate any types of threats in which the knowledge or modus operandi of seen insider threats been acknowledged. The common features which been considered are broken connection based on physical proximity, unresolved alarms threat and interruption between machine by previous researcher [10, 13, 19].

1.2 Related Research in an Insider Threats

This section will discuss about three possible factors lead to insider threats known as human factors, system factors and machine factors. Each factor has a variety of strategies in reviewing the probability of insider threats in order to produce a solution based on the studies undertaken. A number of reviewed articles which fall under the above-mentioned factors categories as one of insider threat contributor been highlighted.

1.2.1 Insider Threats Cause by Human Factor

The human factor can be defined as a human behavior that behaves abnormally which contribute to an insider threats action. For instance, author [4] proposed approaches of differentiating risk-tolerant individuals from risk averse individuals which enable an organization to increase or maintain productivity which can produce summarization of a preliminary set of mitigation strategies and countermeasures of threats. Besides that, author [4] also describe unintentional insider threats which defined as a current or former employee, contractor, or business partner who has or had authorized access to an organization's network, system, or data and who, through action or inaction without malicious intent, unwittingly causes harm or substantially increases the probability of future serious harm to the confidentiality, integrity, or availability of the organization's resources [4]. Besides author above, author [7] proposed the taxonomies analysis that can contribute to the association and disambiguation of insider threat incidents as the protection solution against them. The objective of this approach is to systemize the knowledge gathered in insider threat research but at the same time leveraging current stranded theory method for severe literature review [6]. Moreover, author [9] has derived on how the analyst proposed a relationship between Dark Triad personality traits, related constructs and external process experiences that derived from past collected works using formal modelling methodology. The proposed method Capability Means Opportunity (CMO) model has novel concept that fills in the breach in the image of literature of insider threats personality that can capture the threat capability and opportunity. In addition, author [1] has objectives to deliver an essential consideration for mitigating with insider threats by using technique of conformist insider mitigation resolutions in isolated clouds. The methods are by classifying five dimensions include cloud deployment, source of the threat, threat impact, insider threat approach and susceptible cloud service. Moreover, the experiment and observation lead to the resulting of a different taxonomy of insider threats for the cloud background structures and benefits in the decrease misunderstanding and overlying between classifications, explanation of the necessary perceptions of insider threats and cloud computing [1].

1.2.2 Insider Threats Cause by System Factor

System factor can lead to insider threats such as system failure or errors, operational activities or sequences, and so forth. For instances, author [2] consume and analyses heterogeneous tributaries of data in real-time to detect outlines that might have different tributaries. The proposed system used to detect malicious insider actions in an organization that based on automatically identified unusual behavior associated with different user within the computational network called as RADISH [2]. Besides author above, author [18] proposed system called XABA is a zero knowledge anomaly based behavioral analysis method to detect insider threats through behavior analysis method that learns from context and detects multiple types of insider threats from raw logs and network traffic in real time. The main objective of the system is to produce user behavior profiles, distinguish exclusive behaviors and potential access indicators, and detect meaningful violation patterns based on behavior of user profiles [18]. Moreover, author [8] propose approach used Corporate Insider Threat Detection (CITD) system that are capable of gaining an inclusive feature set that characterizes the user's current activity inside the organization amongst various observations at former time steps and amongst several users. The comparison between extensive ranges of different metrics is to evaluate the amount of anomaly that was displayed through each of them [8].

1.2.3 Insider Threats Cause by Machine Factor

Apart from human and system factors, machine factor also needs to be considered. An author [17] has proposed new defense mechanism from front end of line (FEOL) integrated circuit and back end of line (BEOL) implementation in the equipment to outsource the previous manufacture method that using simple yet does not have updated security features which can gives chances to an insider to launch threat. Besides, author [17] have goals to ensure that the improved proposal has a high error rate by using technique of geometric configuration corresponding technique and machine learning established technique. Moreover, author [13] has stated that insider threat arises as soon as an authorized worker misuse the approvals and carries terrible damages by using legitimate control given to them. Therefore, author [13] used Supervision Control and Data Acquisition (SCADA) approach and proposed method called as SADM or statistical anomaly detection method in electric power of SCADA system. Besides author above, author [19] proposed method to provide with the improvement of machine used by proposed model based TCP addresses Modbus protocol encapsulated within TCP/IP and middleware level detection of IDS. The detection approach can contribute to identify the normal performance of the network traffic flow and can be comparing with SCADA movement against both models to distinguish possible anomalous behavior.

Example of Real Scenario in MES Relate to a Possible Insider Threats
In this section, a real scenario involves in production system created and categorize according to related factors as in Fig. 1. The study of contributing factors of possible insider threats that can be relate in manufacturing fields has been divided into human factors, system factors, and machine factors accordingly.

The scenario designed above called as automated manufacturing execution system control consisting to major phases which is tracking and ordering system in which the data inserted into server (as in Step 1) before proceed to car production system. As illustrated on the first step, the admin will release the related data throughout the system which have connection that link to the database (SERVER). Next, the database server will feed in the data to application server (as in Step 2), once the instruction from human (admin) via the system application is received (as in Step 1). The application server will analyze and deliver the related data to related production shop such as engine, stamping, welding, painting or trimming throughout the switch (as in Step 3). The switch normally employed to send the data to the related destination as there is a number of production Master Programmable Logic Control (MPLC) geographically distributed (as in Step 4). A bunch of data will be stored into MPLC register known as buffering as storage medium. The data will be delivering to the rightful machine in an order of

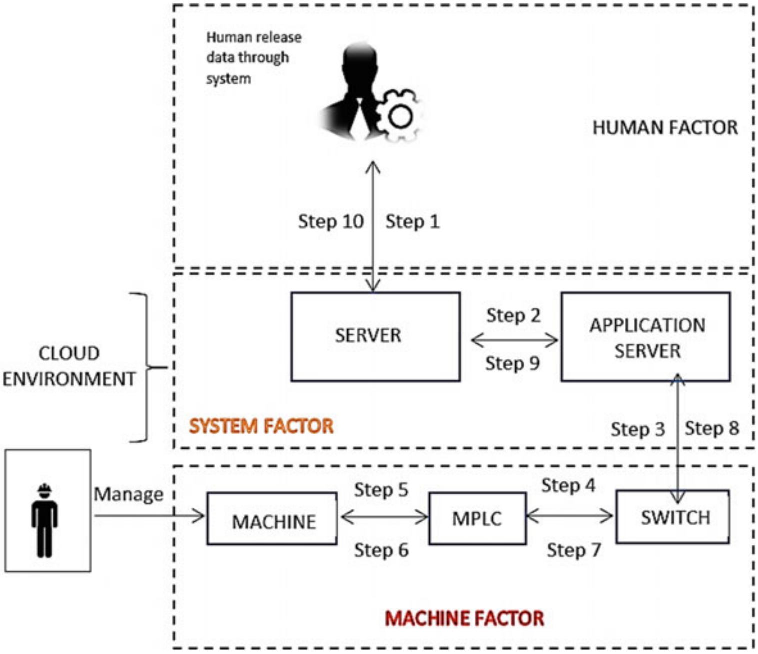


Fig. 1 Insider threat categorization in manufacturing execution system

first-in-first-out (FIFO) manner. The buffer will be updated with new data from the server frequently which based on n-1 concept. For example, if the buffer consisting 10 data of N, whenever the data reduce to 9 (10-1), the MPLC will request new data from the server. The data delivered by the SWITCH in string form unable to understand by the machine. Therefore, this information will be translating by the MPLC into readable form i.e. 0 and 1 which next orders the machine (as in Step 5) to conduct the production. Upon completion of production by the machine, there must be cycle to notify/feedback shows the job has done as well as request for new data. As such, the machine will send completion signal to the MPLC (as in Step 6) and at the same time request for new data (as in Step 5). The completed task feedback will further escalate to the DB Server from the MPLC through Switch (as in Step 6 and Step 8) and Application Server (as in Step 9). The admin can check these completions of activity for tracking purpose throughout the tracking system as shows in (as in Step 10).

Based on above scenario, the human factor insider threats might happen in any steps or phases as claimed by Maasberg et al. [9] in which every person whose act as insider must acquire three main conditions which are motivation, opportunity and capability that can support them to launch threats completely with variable steps. For example, as to relate to above scenario, the employee i.e. admin or line worker could be feeling stress [12] and as a consequence the employee might perform threats (from Step 1 to Step 10) by disrupting the smoothness of process in manufacturing production systems. Moreover, if the worker does not have enough sleep or rest, he/she can be emotionally sensitive [9] and can be less effective during work performance which then lead to make mistakes. The situation can become worst when employee cannot tolerate with working pressure [6] in manufacturing production fields that provide worker with low financial support or monthly payment. Moreover, unintentionally insert wrong data [7] into server (as in Step 1) by the person who have authority can cause wrong data uploaded to application server (as in Step 2) and MPLC (as in Step 4) and finally result in wrong production or machine error. Therefore, from scenario shown above, those human factors that have been summarized can be expected to happen from the Step 1 until the Step 10.

Beside human factors, these studies also have been focused in identifying features of system factor. As mentioned previously, the system factors usually involve an abnormal process of instruction which may cause an impact on automated production systems. There is numerous variant of feature has been discover throughout these study such as employee might have conceptualizing problem on how the production works which involve operational procedure [8] with the system that might resulting to documentation or production failure [20]. In addition, another possible threat which can take place is when the system malfunction which involves occurrence of abnormal system behavior. For example, if the production system faces some technical problem such as incorrect categorization of unique number of car body deliver from database to application server (as in Step 2) supposedly, but the data jump or skip application server and go directly to switch (as in Step 3). This show how the data reached at the switch (as in Step 3) from server without being analyze by application server (as in Step 2) before being distributed to the related shop.

Apart of system factor, some previous researcher also considered machine factors as one of the contributor that can have possibility to become threats such as author [17] describe an organization particularly manufacturing based need to use updated and latest version of machine to produce more secure and effective working environment. The useful of latest version of machine with latest security features should take into account as unfollowing such criteria can grant the chances or probability for the insider to have full knowledge of the machine' vulnerabilities [17]. Once the threat launched, it may cause process of data flow or delivery of each stages disrupted [19] and may give wrong data to production system that could affect the performance of organization output. Furthermore, the tracking and ordering system starts from the beginning of data insertion (as in Step 1) by human (admin) until to the respond or feedback cycle back to the database (as in Step 9), the employee or insider can interrupt that process so that the right data will not being deliver back to the server. This situation can directly affect the performance and reputation of the organization. Possible attempt by insider to shutdown machine is disable the MPLC in order to cut off the receiving data connection from server and update feedback or respond back to the server as illustrated in scenario above, the production system are not able to perform step 3, 4, 7 and 8 due to the interruption [19] in the process by insider. Other than that, during the study also we can assume the machine failure such as power breakdown and mechanical equipment or component failure might also happen during the production system.

1.3 Conclusion and Recommendation

The approach proposed above are some of the mitigation steps to face insider threats as it can have contributed to maintain the confidentiality, integrity and availability of an organizations. As insider threats might not be covered fully by any kind of mitigation, therefore each of the person in an organization or manufacturing should know what their responsibility towards the place they work, how to act as their privilege, why sensitive data should be protect and when should the policy of company being applied. Therefore, recommendation on more effective ways to detect and overcome insider threats especially in manufacturing must be continue. Solid and pure information gather from the manufacturing itself will give huge contribution in order to faces insider threat. There should be an effective and newest detection method for identified errors or threats produce by insiders.

References

1. Alhanahnah MJ, Jhumka A, Alouneh S (2016) A multidimension taxonomy of insider threats in cloud computing. Comput J 59(11):1612–1622. https://doi.org/10.1093/comjnl/bxw020
2. Brock B (2017) Detecting insider threats using radish: a system for real-time anomaly detection in heterogeneous data streams, pp 1–12
3. Elmrabit N, Yang S-H, Yang L (2015) Insider threats in information security categories and approaches. In 2015 21st International Conference on automation and computing (ICAC). IEEE, (ed), p pp 1–6. https://doi.org/10.1109/IConAC.2015.7313979
4. Greitzer FL et al (2012) Identifying at-risk employees: modeling psychosocial precursors of potential insider threats. In 2012 47 th Hawaii International conference on system sciences. IEEE, pp 2392–2401. https://doi.org/10.1109/HICSS.2012.309
5. Greitzer FL et al (2014) Unintentional insider threat: contributing factors, observables, and mitigation strategies, In 2014 47th Hawaii International Conference on System Sciences. IEEE, pp 2025–2034. https://doi.org/10.1109/HICSS.2014.256
6. Homoliak I et al (2018) Insight into insiders: a survey of insider threat taxonomies, analysis, modeling, and countermeasures
7. Homoliak I et al (2019) Insight into insiders and IT, ACM Comput Surv 52(2):1–40. https://doi.org/10.1145/3303771
8. Legg PA et al (2015) Caught in the act of an insider attack: detection and assessment of insider threat. In 2015 IEEE International Symposium on Technologies for Homeland Security (HST). IEEE, pp 1–6. https://doi.org/10.1109/THS.2015.7446229
9. Maasberg M, Warren J, Beebe NL (2015) The dark side of the insider: detecting the insider threat through examination of dark triad personality traits. In 2015 48th Hawaii International Conference on System Sciences. IEEE, pp 3518–3526. https://doi.org/10.1109/HICSS.2015.423

10. Magana J et al (2017) Are proximity attacks a threat to the security of split manufacturing of integrated circuits? IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 25 (12), pp 3406–3419. https://doi.org/10.1109/TVLSI.2017.2748018
11. May CR et al (2017) Insight into insiders: a survey of insider threat taxonomies, analysis, modeling, and countermeasures
12. Moore AP et al (2011) A preliminary model of insider theft of intellectual property. JoWUA, 2(1), pp 28–49. https://doi.org/10.22667/JOWUA.2011.03.31.028
13. Nasr PM, Varjani AY (2014) Alarm based anomaly detection of insider attacks in SCADA system. In 2014 Smart Grid Conference (SGC). IEEE, pp 1–6. https://doi.org/10.1109/SGC.2014.7090881
14. Ntalampiras S, Soupionis Y, Giannopoulos G (2015) A fault diagnosis system for interdependent critical infrastructures based on HMMs. Reliability Engineering & System Safety, 138, pp 73–81. https://doi.org/10.1016/j.ress.2015.01.024
15. Permissions F (2016) A multidimension taxonomy of insider threats in cloud computing
16. Soupionis Y, Ntalampiras S, Giannopoulos G (2016) Faults and cyber attacks detection in critical infrastructures, pp 283–289. https://doi.org/10.1007/978-3-319-31664-2_29
17. Wang Y et al (2016) Front-end-of-line attacks in split manufacturing
18. Zargar A, Nowroozi A, Jalili R (2016) XABA: a zero-knowledge anomaly-based behavioral analysis method to detect insider threats. In 2016 13th International Iranian Society of Cryptology Conference on Information Security and Cryptology (ISCISC). IEEE, pp 26–31. https://doi.org/10.1109/ISCISC.2016.7736447
19. Zhu B, Sastry S (2010) SCADA-specific intrusion detection/prevention systems: a survey an taxonomy, pp 1–16
20. Zou, B. et al. (2018) Insider threats of physical protection systems in nuclear power plants: prevention and evaluation. Progress in Nuclear Energy 104, pp 8–15. https://doi.org/10.1016/j.pnucene.2017.08.006