Studies in Computational Intelligence 540

Zeashan H. Khan
A. B. M. Shawkat Ali
Zahid Riaz *Editors*

Computational Intelligence for Decision Support in Cyber-Physical Systems

Springer

Studies in Computational Intelligence

Volume 540

Series editor

Janusz Kacprzyk, Warsaw, Poland

For further volumes:
http://www.springer.com/series/7092

About this Series

The series "Studies in Computational Intelligence" (SCI) publishes new developments and advances in the various areas of computational intelligence—quickly and with a high quality. The intent is to cover the theory, applications, and design methods of computational intelligence, as embedded in the fields of engineering, computer science, physics and life sciences, as well as the methodologies behind them. The series contains monographs, lecture notes and edited volumes in computational intelligence spanning the areas of neural networks, connectionist systems, genetic algorithms, evolutionary computation, artificial intelligence, cellular automata, self-organizing systems, soft computing, fuzzy systems, and hybrid intelligent systems. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution, which enable both wide and rapid dissemination of research output.

Zeashan H. Khan · A. B. M. Shawkat Ali
Zahid Riaz
Editors

Computational Intelligence for Decision Support in Cyber-Physical Systems

Springer

Editors
Zeashan H. Khan
National University of Science
and Technology (NUST)
Islamabad
Pakistan

Zahid Riaz
Fraunhofer Institute
Bonn
Germany

A. B. M. Shawkat Ali
School of Engineering and Technology
CQ University
North Rockhampton, QLD
Australia

ISSN 1860-949X
ISSN 1860-9503 (electronic)
ISBN 978-981-4585-35-4
ISBN 978-981-4585-36-1 (eBook)
DOI 10.1007/978-981-4585-36-1
Springer Singapore Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014931871

Printed on acid-free paper

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

Preface

Multidisciplinary research has evolved in many innovative areas of science and technological revolution by providing new dimensions as well as challenges in healthcare, robotics, energy, and transportation. Cyber-physical systems (CPS) are no exception to this revolutionary change and many application specific subclasses have already been defined. Examples include aerospace CPS, transportation CPS, healthcare CPS, etc. Cyber-physical systems (CPS) are being developed that are part of a globally networked (with each other as well as with other systems via internet) future world, in which products, equipment, and objects interact with embedded hardware and software. Moreover, when it comes to adding intelligence in CPS by, for example, multiobjective constrained decision algorithms, the conventional CPS merges with the Computational Intelligence for real applications. This has presented some challenging problems, both theoretical and practical, but recently, researchers have witnessed the emergence of some significant results, especially in respect of reducing the computational complexity of high performance computational intelligence algorithms when implemented in low power embedded CPS devices guaranteeing real-time decision support efficiently. This aspect of Computational intelligence for CPS has got little exposure as compared to standalone CPS and CI discussion. This gap is targeted in this volume.

This book is aimed as a ready reference for graduate students and researchers for their knowledge buildup in the application areas of computational intelligence in view of various types of Cyber Physical Systems.

The book has been divided into five parts with respect to CPS application areas. These include wireless sensor and actuator networks (WSANs), Health care and medicine, Robotics, Power and Energy, and industrial applications where we have combined various specialized areas, e.g., logistic systems, aerospace, quality inspection/pre-evaluation, etc.

Finally, we enjoyed going through these impressive chapters which do provide new ideas and scientific research hints, each one within its specific field of study. We wish success for all authors who are keen to perform scientific research in the hope to implement innovative ideas and reliable solutions leading towards a better future for people worldwide.

Zeashan H. Khan
A. B. M. Shawkat Ali
Zahid Riaz

Contents

Part I
Wireless Sensor and Actuator Networks

Configurable, Energy-Efficient, Application- and Channel-Aware Middleware Approaches for Cyber-Physical Systems

Khalid Nawaz, Ilia Petrov and Alejandro P. Buchmann

Abstract The use of Wireless Sensor and Actuator Networks (WSAN) as an enabling technology for Cyber-Physical Systems has increased significantly in recent past. The challenges that arise in different application areas of Cyber-Physical Systems, in general, and in WSAN in particular, are getting the attention of academia and industry both. Since reliability issues for message delivery in wireless communication are of critical importance for certain safety related applications, it is one of the areas that has received significant focus in the research community. Additionally, the diverse needs of different applications put different demands on the lower layers in the protocol stack, thus necessitating such mechanisms in place in the lower layers which enable them to dynamically adapt. Another major issue in the realization of networked wirelessly communicating cyber-physical systems, in general, and WSAN, in particular, is the lack of approaches that tackle the reliability, configurability and application awareness issues together. One could consider tackling these issues in isolation. However, the interplay between these issues create such challenges that make the application developers spend more time on meeting these challenges, and that too not in very optimal ways, than spending their time on solving the problems related to the application being developed. Starting from some fundamental concepts, general issues and problems in cyber-physical systems, this chapter discusses such issues like energy-efficiency, application and channel-awareness for networked wirelessly communicating cyber-physical systems. Additionally, the chapter describes a middleware approach called CEACH, which is an acronym for Configurable, Energy-efficient, Application- and Channel-aware Clustering based middleware

K. Nawaz (✉) · I. Petrov · A. P. Buchmann
Department of Computer Science, TU Darmstadt, Darmstadt, Germany
e-mail: khalid@dvs.tu-darmstadt.de

I. Petrov
e-mail: ilia.petrov@dvs.tu-darmstadt.de

A. P. Buchmann
e-mail: buchmann@dvs.tu-darmstadt.de

Z. H. Khan et al. (eds.), *Computational Intelligence for Decision Support in Cyber-Physical Systems*, Studies in Computational Intelligence 540,
DOI: 10.1007/978-981-4585-36-1_1,

service for cyber-physical systems. The state-of-the art in the area of cyber-physical systems with a special focus on communication reliability, configurability, application- and channel-awareness is described in the chapter. The chapter also describes how these features have been considered in the CEACH approach. Important node level and network level characteristics and their significance vis-à-vis the design of applications for cyber physical systems is also discussed. The issue of adaptively controlling the impact of these factors vis-à-vis the application demands and network conditions is also discussed. The chapter also includes a description of Fuzzy-CEACH which is an extension of CEACH middleware service and which uses fuzzy logic principles. The fuzzy descriptors used in different stages of Fuzzy-CEACH have also been described. The fuzzy inference engine used in the Fuzzy-CEACH cluster head election process is described in detail. The Rule-Bases used by fuzzy inference engine in different stages of Fuzzy-CEACH is also included to show an insightful description of the protocol. The chapter also discusses in detail the experimental results validating the authenticity of the presented concepts in the CEACH approach. The applicability of the CEACH middleware service in different application scenarios in the domain of cyber-physical systems is also discussed. The chapter concludes by shedding light on the Publish-Subscribe mechanisms in distributed event-based systems and showing how they can make use of the CEACH middleware to reliably communicate detected events to the event-consumers or the actuators if the WSAN is modeled as a distributed event-based system.

1 Introduction

Wireless Sensor and actuator Networks (WSAN) fulfill the criteria to be termed as an enabling technology for cyber-physical systems. They enable much richer measurement and control of physical processes than what was possible with wired systems. Cyber-Physical Systems represent a class of systems that are composed of computing devices that monitor and control the real world physical processes. The monitoring task requires these devices to be equipped with sensing elements which provide primary input in the form of raw data to the computing elements of the system. The output of the computing elements is generally channeled to the actuation elements of the system for the desired actuation to take place. The actuation serves as the controlling mechanism for the monitored physical process and also closes the sensing-processing-actuation loop. The interaction of these systems with the physical processes introduces some challenges regarding the physical characteristics such as shape, size and robustness of the devices in addition to the more challenging problem of impedance mismatch between the inherently concurrent physical processes and inherently sequential computing processes. In order for these systems to perform monitoring and control functions on the physical processes, networking of the computing elements, generally on an

ad hoc basis, is also necessary. In a nutshell we can describe cyber-physical systems as being composed of networked, embedded, computing elements that are equipped with sensing and actuation capabilities so that they can monitor and control physical real-world processes.

In Fig. 1 a simplified schematic of a typical cyber-physical system indicating the information flows between sensing, processing and actuation parts is shown. It also shows two optional user interface elements, one is used for configuring the devices and the other one to present the end users with their desired information. Such user interface components mostly find use in intelligent industrial automation systems.

Sensor networks are an essential part of the cyber-physical systems, as they provide necessary input to these systems. This input is processed and analyzed and it could result in a possible actuation, thus impacting some physical, real-world process. Though wired sensor networks have been around for decades, their modern incarnation, wireless sensor networks, was made possible only by the advances in wireless communication in 1990s. Therefore, the sensor networks also went through a transformation due to the progress in technology. Now wireless sensor networks are gradually finding their way into those application areas where wired sensor networks were always used in the past.

Because of the flexibility that wireless communication offers, wireless sensor networks have removed many barriers and have made deployments possible in such scenarios where wired sensor networks were impractical or were very expensive to deploy, e.g. wildfire monitoring with proper coverage, search and rescue in a disaster area, volcanic activity monitoring etc.

Most wireless sensor network deployments collect data from an area of interest called a sensing field. These collected data have traditionally been processed at a central base-station or a sink node. Such a central base-station generally works as a gateway to some other network, proprietary or public, or an actuator. Therefore, in many cases it possesses interface(s) to some other network(s), like the Internet or a private network. In case, some actuators are to be controlled via the central base-station, the collected data is analyzed and relevant actuation decisions are made and then executed by sending a signal to an attached actuator.

It is this model which facilitates intelligent decision making at a central location, i.e. a base-station. Later models diverged from centralized base-station model and focused on distributed processing of sensed data in the network. In-network processing of sensor data has several advantages. One very important advantage is the reduction in communication costs as transport of the raw sensor data, in much larger amounts, to the base station is avoided. In the in-network processing model, the data is processed in the network and only the results are sent over to the base-station. Since processing takes much less energy than data communication in a wireless sensor network, the life of the network can be significantly prolonged by in-network processing of data [21].

The important considerations for such distributed processing of sensor data inside the network and intelligent decision making have many important consequences. The first and foremost of these issues is the selection of processing

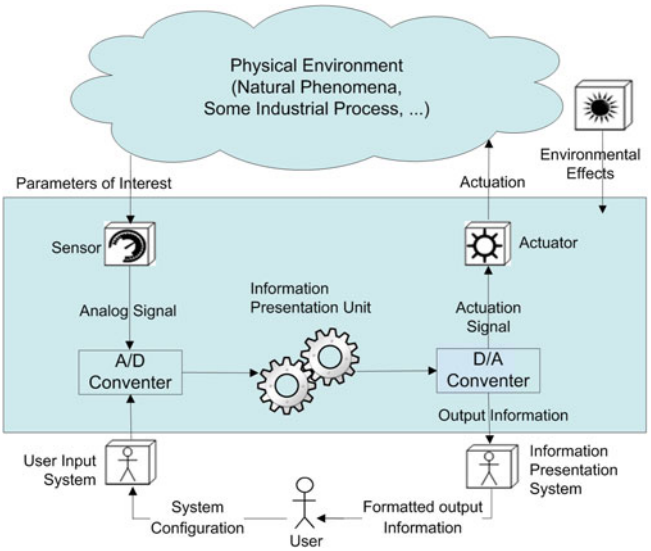


Fig. 1 Schematic of a typical cyber-physical system

centers in the network, which would serve as distributed base-stations, but, in reality they are common sensor nodes. The selection criterion of these sensor nodes should take the application needs as well as the underlying network conditions into account. Since both of these factors vary over time, keeping account of them is not trivial.

The factors that define the state of the underlying network could be divided into two categories, namely, node centric factors and the network centric factors. The node-centric factors characterize a single node in the network. These are the features of a node that influence the role of a node in the network. Examples of the node centric features are a node's residual energy, location vis-à-vis the base-station, hardware capabilities, communication range, and communication power variability, message servicing/processing rate etc.

On the other hand the relationship that binds the nodes into a network starts at a node's immediate neighborhood. The factors that termed as network centric are the ones that determine a node's significance in the network. Examples of Network centric factors are: Size of a node's neighborhood, the strength of the communication links between the node and its direct neighbors, the set of down-stream direct neighbors of a node, the set of upstream neighbors of a node, and network density (number of nodes per unit area in the sensing field).

The interdependence between some of the network centric and node centric factors is also noteworthy. For instance, the size of a node's direct neighborhood is determined by the radio output power chosen by the node. And in fact, the size of a node's direct neighborhood is strongly connected with the node's capability to transmit radio signals. The variability in the transmission power varies the number

of nodes which can listen to the transmission and can intelligently decipher it thus determining the size of the direct neighborhood of a node.

Having such features in a middleware service which can provide configurable and energy-efficient solution keeping the demands of different applications in perspective is an idea worth considering. Additionally, the unreliable nature of wireless communication also necessitates some reliability mechanisms in place otherwise achieving reliable control of physical processes in a cyber-physical system will remain an elusive goal.

Like many other promising research areas, there are several issues in cyber-physical systems that arise with wireless sensing and actuation which are also topics of active research in the research community today. Some of these issues are:

- Appropriate data models to represent the distributed sensor data.
- In-network versus Base Station (centralized) processing of sensor data, thus influencing the actuation model as well.
- Impedance mismatch between the application-layer and the current middleware approaches for cyber-physical systems.
- Reliable communication between participating nodes in the network.
- Appropriate Wireless Networking Standards from a cyber-physical systems' perspective [39] which can withstand interference from other co-existing wireless communication systems.
- Localization of sensors and actuators and suitable abstractions for information exchange between them.
- Time synchronization issues which enable coordinated sleep and wakeup times thus allowing for intelligent use of limited energy resources that are usually available to these systems, and also enable time-stamping of sensor data with globally meaningful time stamps to assist in coordinated actuation.
- Resource constrained nature of wireless sensor and actuator networks in terms of computing, memory, communication bandwidth and limited battery power is also a major concern.

1.1 Application Areas

There are numerous areas where cyber-physical systems can be applied. The CeNSE Project, Central Nervous System of the Earth [35], by Hewlett Packard is an example of an ambitious large scale cyber-physical system. The aim of the project is to revolutionize the way information is gathered, processed, analyzed and acted upon, i.e. actuation. The planned system is envisaged to have billions of nanoscale nodes embedded with sensors that can feel, smell, taste, see, and hear the environment around them and report it back to the core of the central nervous system which will be composed of powerful computing engines. These computing engines are envisaged to have so much computing power that they can analyze and

act upon the huge volumes of sensory data in real-time. The actuation part is envisaged to consist of new breed of business applications and web services. How effective this system is going to be in achieving its stated goals remains to be seen?

Additionally, the applicability of cyber-physical systems in controlling industrial processes in intelligent industrial automation is also a promising prospect for future and can change the way industrial automation is done today. The issue of reliability in wireless communication, however, is a major unresolved problem in industrial communication systems. Moreover, the cycle-times in industrial bus-systems are much shorter than what could be achieved with ZigBee, 802.15.4 based networks or other wireless communication standards which are prevalent in cyber-physical systems today.

Additionally, cyber-physical systems have been successfully applied in scenarios like habitat monitoring [25], precision agriculture [18], military applications like surveillance and reconnaissance [1], health applications like patient vital signs monitoring and emergency response [28], smart home applications [34], industrial applications like structural health monitoring [19], elderly care [37], and for ensuring personnel safety in industrial setups [30], automotive systems [38], advanced electric power grid [27, 32], Vehicular Ad-hoc Networks (VANET) [43] etc (see Fig. 2).

1.1.1 Industrial Process Control

Cyber-physical systems find use in Industrial automation, especially at the lowest end of the automation spectrum where sensors and actuators interact with the physical world. Actuator-Sensor Interface (*AS-i*) [2] and similar standards define the networking of the sensors and actuators to form a cyber-physical system for industrial automation and control. The current *AS-i* standard, version 3.0, and all previous versions of it define a cable-based network. The current state of the wireless communications technology, especially in the low-power and low bandwidth range, has reliability issues and is not mature enough for safety–critical applications in the industrial automation world. Another constraint to the adoption of wireless technology in the safety–critical industrial automation systems is that the cycle times for data exchange, in safety–critical industrial automation applications, are also very short. For instance, the *AS-i* standard mandates that the cycle time be 5 ms or less for an *AS-i* network consisting of an *AS-i* Master and up to 62 *AS-i* slaves, both sensors and actuators. During this cycle time the *AS-i* master should be able to query all the connected *AS-i* slaves. These cycle times are difficult, if not completely impractical, to keep if low-power and low-bandwidth wireless communication is used.

That said, there are successful adoptions of wireless communication, though at higher bandwidths and with high power radios, in industrial applications also. Industrial Wireless LAN (IWLAN) could be cited as an example in this regard. IWLAN uses IEEE 802.11a/b/g/n standard based radios to provide wireless

Fig. 2 Different application areas of cyber-physical systems

connectivity at higher transmission rates (up to 450 Mbps). Such wireless data transmission rates are suitable for video streaming from video sensors and could also be used in surveillance applications. IWLAN offers good prospects for a possible wireless *AS-i* standard in future, since the available bandwidth allows multi-channel simultaneous data exchange with all the connected slaves within the cycle time of 5 ms.

1.1.2 Underground Mining Applications

Underground mining represents a good application scenario for cyber-physical systems. The need to apply environmental controls inside an underground mine is well known in the mining industry. The health and safety of miners is directly

coupled with the success of these environmental control measures. With regard to the safety of mine workers, control of the underground environment by using actuators like ventilators for bringing in fresh air and expelling the hazardous gases out is very important. Having a network of sensors deployed inside a mine and the actuators like ventilators controlling the mine environment, offers a good practical solution. Nawaz et al. [31] explain an underground mining scenario for the application of wireless sensor and actuator networks with a novel concept of proximate environment monitoring using intelligent mobile agents. The writers extend the same concept to other industrial set-ups, for example, chemical processing plants where there is always a danger of having some industrial accident and leaking of hazardous substances into the air making it unsafe for the workers.

1.1.3 Participatory Sensing Applications

The widespread use of smart phones with embedded sensors has given rise to a different kind of cyber-physical system based upon the principle which is popularly known as participatory sensing. In their seminal paper on the subject [13] explain how a citizen powered approach like participatory sensing works. Their envisaged participatory sensing system is depicted in Fig. 3.

According to the authors a participatory sensing system works on the principle of data collection/sensing by a group, small to very large, of people, using smart devices with embedded sensors and wireless communication capabilities, e.g. smart phones, doing routine activities and sharing these data ubiquitously through cellular networks. These data are analyzed and processed in the cloud. People interested in these data can query the Cloud which provides them with the results of the analysis done on raw data in the form of simple to understand visualizations. Lau et al. [22] suggest a participatory sensing based cyber-physical system in the public transport domain. Their prototypical system is named ContriSenseCloud and follows client–server model of computing and assists people in sharing and querying contributed data by the people.

A participatory sensing system, in principle, is a system that is powered by the people, for the people and the participation of people in it is purely on a voluntary basis. People can learn through experiences of others using participatory sensing and can then coordinate and plan their activities accordingly. This informed coordination and planning of activities using analysis data obtained from the participatory sensing system can be regarded as the actuation part in a cyber-physical system context.

The early trends indicate that smart-phones are becoming popular than ever and if the privacy concerns are addressed properly the use of participatory sensing will continue to rise at a rapid pace in more interesting scenarios of human activity.

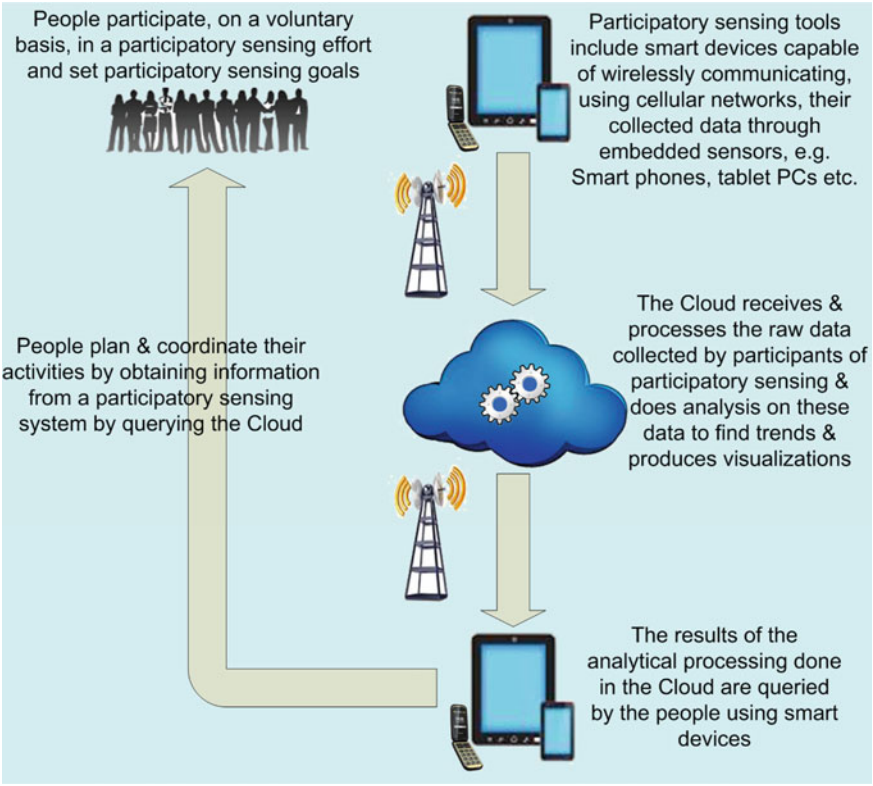


Fig. 3 A system based upon the principle of participatory sensing as envisaged by Goldman et al. [13]

1.1.4 Fuzzy Logic Based Approaches in Cyber-Physical Systems

The term fuzzy logic was first coined in the seminal work done by Zadeh [42] on fuzzy set theory. Since then fuzzy logic has been applied to solving problems in many systems including control systems modeling problems, because it can act as a bridge between qualitative and quantitative modeling of such systems. Since many cyber-physical systems are in fact control systems, the application of fuzzy logic based models for solving cyber-physical systems' problems is a natural consequence.

Fuzzy logic is based upon the principle of mimicking the human ability to reason and arrive at a conclusion or a decision despite the vagueness of input parameters. This reasoning follows a finite number of rules and produces a fuzzy output which is defuzzified into a smooth output often using interpolation. The rules are based upon multi-valued logic, instead of classical logic where we have absolute truth and false values for variables. A fuzzy logic variable could belong to multiple fuzzy sets to a different extent or degree, i.e. the degree of membership of

fuzzy logic variables to fuzzy sets varies or partial memberships of fuzzy logic variables to fuzzy sets is possible.

Following the fuzzy set theory presented by Zadeh, a fuzzy subset X of a crisp set Y is obtained by assigning to each element x of X the degree of membership of x in Y. A very simple example of fuzzy sets is the following. Consider Y to represent the set of all soccer players in German national football team and X to be the set of "*tall*" and "*good*" soccer players in the German national football team. Since the definition of "*tall*" and "*good*" is qualitative and subjective, according to the fuzzy set theory of Zadeh, we can only assign soccer players to the fuzzy sets based upon their degrees of membership. A given player could be "*tall*" and "*short*" at the same time, to varying degrees of membership, in these fuzzy sets. As the height of a player increases, his membership into the fuzzy set "*tall*" increases and his membership into the fuzzy set "*short*" decreases. The same logic applies to the membership degree of a player into "*good player*" and "*bad player*" fuzzy sets.

Since the antecedent propositions could be compound, as in the soccer player example, one needs to use compositional rules to obtain a numeric value representing the degrees of memberships of the individual conditions in the antecedent proposition to the fuzzy sets. Since the propositions are mostly either conjunctive or disjunctive, commonly used methods for combining them are to take a conjunction (AND, which means taking a minimum) or a disjunction (OR, which means taking a maximum). For instance, if a player has a height of 180 cm and if the number of goals scored by him per game is 0.3, should he be termed as a "*tall*" and a *"good player"*? Considering his height, 180 cm, he might belong to the fuzzy set "*short*" with a membership degree of 0.6 and to the fuzzy set "*tall*" with a membership degree of 0.4. Additionally, the goals scored per game value 0.3 might get him a place in the "*average player*" fuzzy set with a degree of membership of 0.8 and a place in the "*good player*" fuzzy set with a degree of membership of 0.2. Such partial memberships of fuzzy logic variables to fuzzy sets make the fuzzy logic stand apart from the classic logic where the variables could either have a *True* (1) or a *False* (0) value. Further explanation of fuzzy logic is given in a subsequent section of the chapter where it is applied to solving a concrete problem in wireless sensor and actuator networks.

1.2 Important Issues for Cyber-Physical Systems

This section describes some of the issues that are important for cyber-physical systems. These issues are also a topic of research in the community and very promising solutions to these issues are being suggested.

1.2.1 Data Models to Represent Distributed Sensor Data in WSAN

The sensing component of a cyber-physical system is composed of sensors that sense the environment and produce a collection of raw bytes. Similarly, the

actuation part of these systems is composed of actuators that require digital or analog signal. Each sensor node in the network can potentially be equipped with multiple sensors, each producing different data reading. When semantic information is added to these data, it changes these data from raw form to their logical semantic and more meaningful form. The representation of these sensor data is also challenging especially when one considers the ways to visualize, access, process and present these data. Once the semantic information has been added to these data, they remain no more a collection of raw bits and bytes; rather they are enriched with their semantics which make them meaningful for the applications and the end users alike. The representation of these data needs appropriate data models, an aspect which was ignored by many earlier protocols and middleware approaches. That said, there have been some promising efforts in this regard like the Cougar approach [40] and TinyDB [24]. Additionally, the work done by [15] in combining the network abstraction called scopes with the declarative SQL-like queries offered by TickyDB is another promising approach in this regard. Such a representation of sensor data with appropriate data models has made the understanding and visualization of the data easier for the end users as well as for the designers of such systems. Writing simple queries on virtual sensor tables, where each sensor reading forms a column and when combined with the timestamp and location information it forms a row in the table, and associating their processed output to the actuators which may cause required actuation is much simpler than, say, writing applications in programming languages like C. Such approaches have enabled the researchers, in natural sciences and in other disciplines, to use sensor and actuator networks and design and write applications themselves rather than communicate their requirements to some programmers who could write programs for them.

This has also given rise to further research on topics like in-network efficient joins [17] using virtual sensor tables like representation of sensor data and query distribution and processing in the network.

1.2.2 In-network Versus Base-Station Processing of Sensor Data

In order for a cyber-physical system to make intelligent decisions the input comes from a sensor network. Traditionally these sensor networks have had wired communication but with the improvement in technology the micro-electro-mechanical sensing devices equipped with wireless radio modules emerged. These devices when deployed in a field network with each other forming what is commonly known as a wireless sensor network. These sensor networks generally are structured in a way that a base-station or a sink node acts as the root and the rest of the sensors form a routing tree rooted at the sink node. Generally these sensors sense and send raw data to the sink which acts as a gateway providing the actuators with an actuation signal after processing the raw data. However, this model is in-efficient as the sensor nodes deplete their batteries quickly because of sensing and sending raw data without any processing. Later protocols diverged from this

sense and send model and moved to sense and send only if needed model, like if the data exceeds a given threshold. Still these threshold based models were in-efficient as when the threshold exceeds the data is still sent without applying any processing or aggregation scheme. Another improvement over the threshold based model is sense, process and send only if needed approach. This model is popularly known as In-network processing. This model is a good improvement over the other two models and has been able to successfully improve the sensor network lifetime and as a result the lifetime of the whole cyber-physical systems.

In Industrial automation systems the wired sensor and actuator networks are still dominant and in-network processing of sensor data is not the norm. The data is generally processed at a central controller or a master node and the actuators are also controlled from there. Examples of such systems are Actuator-Sensor Interface (ASi), Controller Area Network (CAN) and Controller Area Network Safety (CAN-Safety), CAN in Automation (CiA), CANOpen [5], DeviceNet etc.

1.2.3 Impedance Mismatch Between the Application Layer and the Middleware

One of the primary goals of a middleware is to provide services to the applications that match the needs of applications. Even better strategy for a middleware is to show adaptive behavior in the face of changing needs of the applications under evolving network conditions. However, most middleware for cyber-physical systems fail on these counts. This problem of impedance mismatch between the needs of applications and the services that a middleware offers makes many middleware unsuitable for real-world deployments. Ideally speaking, for application developers focusing on the application logic and its associated problems should be the only concern. Using a middleware that offers little qualitative and adaptive services to the applications make the problem worse for the application developers as they not only face the problem of implementing the application logic but also the problem of taking care of the issues in the layers underneath the application layer while developing applications. The impedance mismatch between the middleware services and the application needs thus adds to the worries of the application developers. Therefore, in order for cyber-physical middleware to avoid this impedance mismatch, the dynamic adaptive behavior in their services should be provided.

1.2.4 Message Transport Reliability in the Network

As an essential part of a cyber-physical system the sensor networks need to provide the input required for intelligent decision making in these systems reliably. The wireless communication has always reliability concerns and it is the single most important factor for keeping this technology out of the most safety related industrial applications. Message retransmission schemes on the transport layer

tend to increase the message delivery costs. The schemes incorporating either explicit message acknowledgements (*ACK*) or implicit acknowledgements (*IACK*) and their variants are widely used to provide better message delivery reliabilities.

1.2.5 Localization of Sensors and Time Synchronization Issues

As opposed to other distributed systems the link between a sensor node and its location is much stronger. The reason for this stronger link is very obvious. In most applications not only the sensed information is of significance but also the place this sensor reading is taken at is equally important. Without the location information the sensor reading will be meaningless. Therefore, the localization of sensors is an important issue in sensor networks and thus in cyber-physical systems also. Determining the sensor locations is not very trivial as most sensor nodes are without any positioning hardware on them, e.g. GPS modules. The GPS modules would make the cost of the nodes higher as well as they would drain the batteries of these nodes quickly. Another problem with the GPS based localization is that the GPS signal doesn't get through dense vegetation and inside buildings. The positioning methods based upon directional antennae have also been suggested but again such special hardware increases the cost of the sensor nodes. Methods based upon triangulation and time of flight and time of arrival of radio signal have shown promise and are mostly used also.

Time synchronization issues in a sensor network are also very important to consider for two reasons. The first and foremost is to maintain synergy between sensing and actuation in a real-time system where late actuation is equivalent to no actuation. The second reason is that the sensor readings make sense only in the context of space, where they were taken, and time, when they were taken. Maintaining the sensor clocks synchronized, by no means, is an easy task. There have been several protocols that have been suggested in the literature for this purpose [33] but most of them have high overhead. Also keeping a protocol's overhead low and still ensuring a good degree of time synchronization is a challenging task.

1.2.6 Hardware Constraints

The nodes used in wireless sensor and actuator networks are known to have constrained resources in terms of memory, processing power, sensing range, communication bandwidth, and communication range. Generally these nodes are battery powered. Therefore, ensuring network longevity by controlling battery usage is of critical importance. Different energy scavenging or harvesting techniques have been applied to enhance the operational life of sensor nodes. However, the additional hardware required for the purpose increases the cost of the nodes to make them infeasible for the applications where low-cost nodes are a prerequisite for the success of the deployment.

The controllers used for process control in industrial automation, like AS-i master nodes that lie at the interface between the more resourceful fieldbus devices and less resourceful actuators and sensors, generally possess better hardware resources than the sensors and actuator elements in the system. That said, these controllers still are considered constrained in resources and whatever software is written for them has to work within these constraints.

1.3 Additional Issues for Cyber-Physical System Middleware

In addition to the issues discussed in the previous section, there are issues that are important for the applications in cyber-physical systems. This section describes some of them. The middleware approaches that are developed for cyber-physical systems should consider these issues important and they should have mechanisms in place to cater for these issues. It is by no means a comprehensive list but it does shed light on important issues like configurability, energy-efficiency, application and channel-awareness.

1.4 Putting Configurability into Context

Configurability in the context of a cyber-physical system middleware is the ability of the middleware components to adapt them to better cater for the needs of the applications running and using their services. This implies that the applications, that use the services provided by an underlying middleware, should be able to dictate their preferences to the latter or a component thereof. The middleware and its components on their part should have the capability to adapt to the needs of the application. One aspect of this adaptability could be to make several protocols available that an application can choose from. However, in resource constrained devices such a connotation of adaptability is difficult to support. Therefore, CEACH protocol, which is developed around the ACDMCP protocol presented in [29] and is described later in this chapter, also provides a range of choices for the applications. It allows applications to dictate their preferences on a number of aspects including, among others, residual energy, link reliability, hop-distance, and neighborhood size of a node in the network.

1.4.1 Energy-Efficiency

Since the sensor nodes used in most sensor network deployments are operated on two AA size batteries, it is one of the primary concerns for network designers as well as for application developers to follow such network architectures as well as application designs that enable conserving the battery power of the nodes as much

as possible. One might wonder why we can't use more powerful batteries that can hold longer and keep the network operations intact for a longer period of time. However, due to form- and cost-factor of these devices, more powerful batteries can not be put on them.

One of the most important criteria for the success of an approach for these systems is its energy efficiency. Since energy efficiency has a direct impact on the life of the network, the longer a node holds on to its battery power, the longer is the network lifetime. Additionally, disconnections in the network, due to the death of critical nodes in the network, are a known problem. It has been observed in most sensor network deployments that the sensor nodes that lie close to the base station die early than their counter-parts which lie further away from the base-station. One logical reason for this phenomenon is the fact that the nodes that lie close to the base-station have to process more messages on average than the nodes that lie further away from the base station. Some in-network processing techniques have shown promise in this regard, since they focus on processing the data in the network and apply data aggregation techniques to lower the network traffic. Therefore, a balanced use of energy in the network is a desired feature in most sensor network protocols.

1.4.2 Application-Awareness

Application-awareness is another desirable feature in cyber-physical system middleware but there are few approaches that actually possess this feature. Application-awareness, in our opinion, can be categorized into two classes, namely, adaptive and non-adaptive or static. Any middleware providing compile time switches to adjust its services for the applications running over it belongs to the non-adaptive or static application-awareness class. On the other hand, the middleware that provides for mechanisms for run-time behavioral changes in its services based upon application demands and the evolving network conditions belongs to the adaptive application-awareness class. It is clear from the definition of these classes that providing for adaptive application-awareness is much more challenging than providing for non-adaptive one.

1.4.3 Channel-Awareness

Wireless links are prone to interference from a range of sources. These interference sources include other co-existing concurrently happening wireless communications in the ISM[1] band in addition to the multi-path interference and the fact that a group of nodes are sharing the same wireless channel. As a result many

[1] The ISM band is the radio band reserved internationally for the use of radio frequency energy for industrial, scientific and medial purposes.

protocols suffer performance problems and have low message delivery ratios. Especially in a cyber-physical system message communication reliability takes a completely different dimension and becomes much more relevant than, say, in an experimental setup or a fun application. If a cyber-physical system is deployed to do some safety–critical actuations based upon sensor input in a distributed network, then certain message transport reliability guarantees have to be given to meet the safety needs of the application.

If a given protocol takes the dynamics of the wireless channel into account and takes measures to make the better use of this information for improving message delivery ratios, we can say that the protocol is channel-aware and this feature itself is termed as channel-awareness.

Unfortunately most application protocol developers leave the provision of message transport reliability on the transport layer in the protocol stack [29], if the transport layer is present at all. Some protocols do implement mechanisms for channel-quality assessments and choose routing paths accordingly. Collection Tree Protocol (CTP) suggested by [12] could be cited as an example in this regard. Therefore, these protocols, which consider link reliabilities in the network layer especially, have higher message delivery ratios than the ones which don't. The experimental results presented later in the chapter attest to this fact.

1.4.4 Hierarchical Clustering

Traditionally message communication in a WSAN or a WSN has been flat. The need to structure these networks gained prominence as the size of the deployments grew and the need to process the data in the network arose. Since in smaller deployments each node could directly send data to the sink which would process it and make intelligent actuation related decisions accordingly, earlier protocol developers didn't feel the need to structure the network into some sort of a hierarchy. However, soon it was realized that even in single hop WSN considerable energy savings could be achieved by clustering the network. Initial efforts to hierarchically structure the network e.g. LEACH [14], represent an effort in this regard. Later efforts departed from the single-hop network model assumed by earlier protocols and clustering protocols for multihop sensor networks began to emerge, e.g. HEED protocol [41].

There are also quite a few clustering protocols that have been suggested for Mobile Ad hoc Networks (MANETS for short). Protocols like LCA [3] and WCA [6] which have been suggested for MANETS have limited applicability in WSN due to their focus on mobility of the nodes than on other attributes which are more relevant in WSN.

The CEACH middleware service, described later in the chapter, utilizes hierarchical clustering in an adaptive and application-aware way to provide for in-network processing.

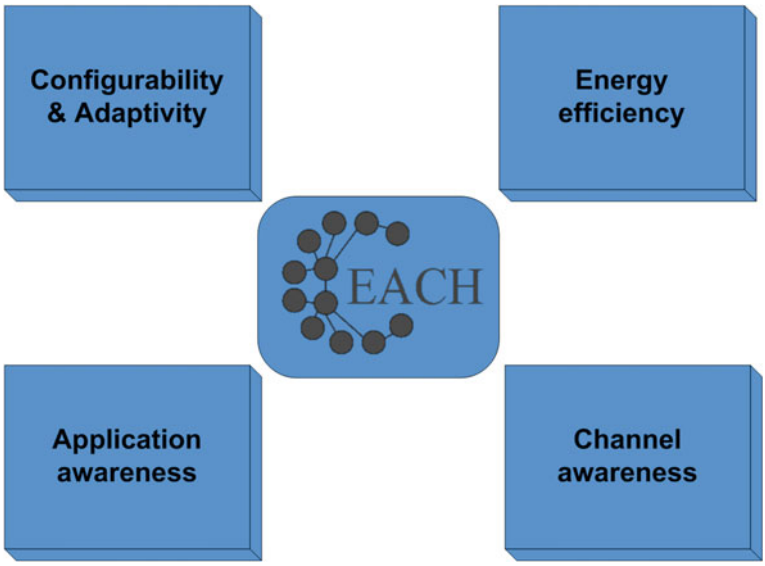


Fig. 4 Major CEACH features

1.5 The CEACH Approach

CEACH is an acronym for Configurable, Energy-efficient, Application- and Channel-aware, Clustering based middleware service for WSAN. This section begins with a description of the major features of it (see Fig. 4). That is followed by the description of the network model assumed by CEACH. Important design goals that are set for CEACH have been included also. The importance of channel awareness and the way it has been accomplished in CEACH and its associated benefits get a mention also. The description of how applications can dictate their requirements to a network layer that models the network into a hierarchical structure is also described. The cost metric used in CEACH for choosing cluster heads distributedly and its configurability is also mentioned.

1.5.1 CEACH Features

Any middleware approach targeted at the resource constrained platforms like WSAN should possess features that make it resource aware. One very important consideration that the designers of many middleware services for WSAN overlook is meeting the needs of the applications in the face of evolving network conditions.

Feature	CEACH	CTP	HEED
Usage as a Middleware Service	+	−	−
Application Awareness	+	−	−
Channel Awareness	+	+	−
Energy Awareness	+	−	+
Adaptive Tx. Power Control	+	−	+/−
Adaptive Beaconing	−	+	−
Any-hop Clustering	+	/	−
Graceful Node Failure Handling	+	+	−
Load Balancing	+	−	−

Fig. 5 Feature comparison of CEACH with CTP and HEED. If an approach possesses a feature it is shown with a '+' sign, if it doesn't it is shown with a '−' sign, and if it is an irrelevant feature, it is shown with a '/' sign

Ignoring these conditions make such solutions vulnerable to excessive message loss. Consequently, these protocols end up showing poor performance in terms of data delivery ratios and reduced network lifetime.

In what follows, some of the features that the CEACH middleware service possesses are described. They are presented comparatively with other relevant approaches like HEED and the Collection Tree Protocol (CTP) [12]. Figure 5 presents the same comparison in a more succinct form.

Usage as a Middleware Service

One of the prime features that CEACH possesses is that it is conceived as a middleware service and not merely as a network layer service which is agnostic to the needs of the applications running on top of it. This is an important feature that makes CEACH stand out among other clustering services like HEED and/or link reliability based collection protocols like CTP. As a middleware service it offers different data reduction functions like *MIN*, *MAX*, *AVERAGE* as well as different data aggregation mechanisms. This enables applications to choose their desired data reduction or a data aggregation function.

CEACH and Application Awareness

Applications running in a distributed WSAN could possibly have very different requirements from one another. Experience has shown that the application developers only like to focus on implementing the core application logic in the application layer and the rest of the services, which are the responsibility of the lower layers in the protocol stack anyway, they want them to be selectable and configurable. The CEACH approach has been developed with these priorities in mind. It provides applications enough freedom to specify their priorities, so that they can achieve the desired shape and behavior of the network with the end result of higher message delivery ratios and higher network lifetime.

Channel Awareness in CEACH

One of the major advantages that CEACH offers over other clustering services is the channel-awareness feature in it. The core clustering service in CEACH incorporates

link reliability mechanisms to improve the operational performance of the network in terms of data delivery ratios and prolonged network lifetime. Unreliability of the Wireless channel is one of the major factors for the non-adoption of this technology in safety critical applications. It is only the channel aware protocols and middleware that offer some promise for such applications. CEACH middleware service looks a promising advancement that has the potential to achieve the same end. Channel quality in CEACH is regarded as a property of the wireless link between two communicating nodes in the network. The links in a wireless network tend to show directional properties and link reliabilities can be different between a pair of communicating nodes in either direction. For the same reason, CEACH considers the wireless links as directional and tries to assess message communication reliability on directional basis.

Energy Awareness in CEACH

Wireless ad-hoc networks like WSN and WSAN have traditionally been composed of energy constrained nodes. The sensor nodes are generally operated on a pair of alkaline (AA) batteries with limited lifetime. Research has shown that the major energy consuming operation in the nodes is the radio. Therefore, for any protocol or a middleware approach to be successful in such networks, it should possess energy-awareness features. Despite this requirement, many approaches disregard this factor altogether and doesn't consider it a protocol design constraint. CEACH middleware service incorporates such mechanisms that ensure better use of the energy resources available to the nodes. Channel-awareness in CEACH also translates into energy-awareness, since it is channel-awareness which enables the nodes choose better cluster heads. Additionally, these cluster heads then choose the best cluster heads as their upstream neighbor for upstream data traffic in the network towards the base-station or the sink node. This happens only if the application scenario demands it. The core clustering service also incorporates residual energy of the nodes in choosing the cluster head nodes. This increases the probability that the network structure will last longer without any reconfigurations.

Some WSN approaches try to conserve node battery power by utilizing efficient duty cycling of the radio [4]. Such approaches work at the MAC layer and have shown promise. However, the designers of the upper layers have mostly left the energy awareness issue for the MAC layer to tackle. CEACH diverges from this trend and applies energy-aware mechanisms in the network layer on top of the duty cycling mechanisms offered by the MAC layer underneath.

Adaptive Transmission Power Control

This is another mechanism that is useful in conserving battery power of the nodes. The basic physics books on radio communication tell that larger the distance between communicating nodes, larger will be the communication cost. Considering a perfect noise and interference free channel, this principle holds true. Therefore, the nodes should not use maximum radio power available to them when they communicate with the nodes lying in close vicinity. However, implementing such mechanisms is not trivial. CEACH approach incorporates adaptive transmission

power control in its core clustering service for both intra-cluster as well as inter-cluster communication.

Any-Hop Clustering

Most clustering approaches like HEED, LEACH and other similar approaches assume 1-hop clusters. The LEACH approach is even restrictive as it assumes that all nodes can directly communicate with the base-station or the sink node. This reduces the flexibility of these approaches especially in such deployments where node density is non-uniform. CEACH approach, on the other hand, offers applications running on top of it to dictate the size of cluster themselves. The feature is called any-hop clustering, since the application can decide the size of clusters dynamically. If the application sees that in a given round of clustering the data delivery ratio was low, despite assigning maximum importance to link reliability feature, it can choose a different size of clusters for the next round of clustering. Since cluster size has an impact on the level of data aggregation that is possible in the network, choosing larger size clusters can improve the data aggregation efficiency of the network.

Adaptive Beaconing

In order for a protocol to maintain its topology and keep its estimate of the link quality updated, it needs to send short probe messages periodically on each of its links. Sending such beacons too often enables the protocol keep its topology well intact as well as a good estimate of link quality within its neighborhood. However, it increases the communication overhead of the protocol, thus impacting the network lifetime negatively. Since the wireless links tend to be dynamic, their quality assessment needs to be continuously updated. The CTP protocol uses a mechanism called adaptive beaconing, which simply means that a node's beacon rate to its neighbors is dependent on the link quality assessment itself. This enables a node to have a low beacon rate, if the network is stable and vice versa.

The link assessment module in the CEACH approach diverges from the beaconing model and uses actual data packets to assess link quality. This approach has two advantages. First it doesn't have any beaconing overhead. Second the link assessments done using short beacons have lower probability of packet collision than, say, actual data packets, simply because the latter are generally longer and, thus, occupy the wireless channel for a longer period of time. Therefore, the link assessments done using short beacons tend to overestimate the link quality.

Graceful Node-Failure Handling

In a static WSN, the most probable reason for a node failure is the battery exhaustion, i.e. nodes generally fail, in a static network, when their batteries run out of charge. Suddenly disappearing nodes add another dimension of complexity to the problem of topology maintenance in the network. The beaconing mechanism, as described in the previous section, can help resolve this issue as well. However, the beaconing mechanism does discover dying nodes on a reactive basis, i.e. if the link quality keeps on degrading on a given link and a node keeps on not

receiving any acknowledgements back for its sent data packets, at certain point in time the node declares the given link and its counterpart node as dead. However, this entails lost data-packets and too many message (re-)transmissions (data, acknowledgements, and beacons) resulting in higher energy consumption.

CEACH utilizes a more pro-active approach in discovering node failures. Each node keeps track of its residual energy and that of its immediate neighbors. This data is exchanged as part of cluster formation protocol. When a node runs low on battery power, it sets energy-bit in its outgoing messages indicating that it is running low on battery power and might soon not be available. Other nodes then start looking at alternate nodes in their neighborhood to keep the network topology intact. Secondly, when the nodes know the energy levels of their neighboring nodes, they already make intelligent decisions while choosing their upstream neighbors on the route towards the base-station. This is an additional assurance that the nodes communication partners possess enough energy to keep the topology intact for longer period of time than, say, in those protocols, like CTP, where residual energy of the nodes is not considered.

Load Balancing

In protocols that follow a greedy approach, like CTP, HEED and many others, in choosing the routing paths, it is generally not possible to balance the message servicing load in the network. The nodes which possess high energy or better link quality generally end up servicing more messages, because they are chosen by more of their neighbors as communication partners thus creating hot-spots in the network. Therefore, these nodes which lie at these hot-spots consume their batteries at a higher rate and tend to die early. This results in a reduced total network lifetime.

In CEACH approach, load balancing mechanisms are also in place which prevent the hot-spot problem. Firstly, CEACH doesn't use a single node or network parameter, like residual energy or link quality, to make topology related decisions. Secondly, the graceful handling of the nodes that exit the network because of running low on battery power by making them inform their neighborhood about it. Thirdly, cluster heads, which have the potential of becoming a natural hot-spot, when consume their batteries up to a threshold give up their cluster head role, preventing the hotspot and an early death problem.

1.5.2 CEACH Architecture

The CEACH middleware is architected in a way to provide services to the application layer directly as well as to any other middleware that requires services from it. Therefore, an application running in the network can dictate its preferences to the CEACH middleware service directly or through another middleware regarding its network structuring priorities. The size of the node groups or clusters, the characteristics of the nodes which should get priority in becoming group leaders or cluster heads, their 1-hop neighborhood size, their residual energy, and

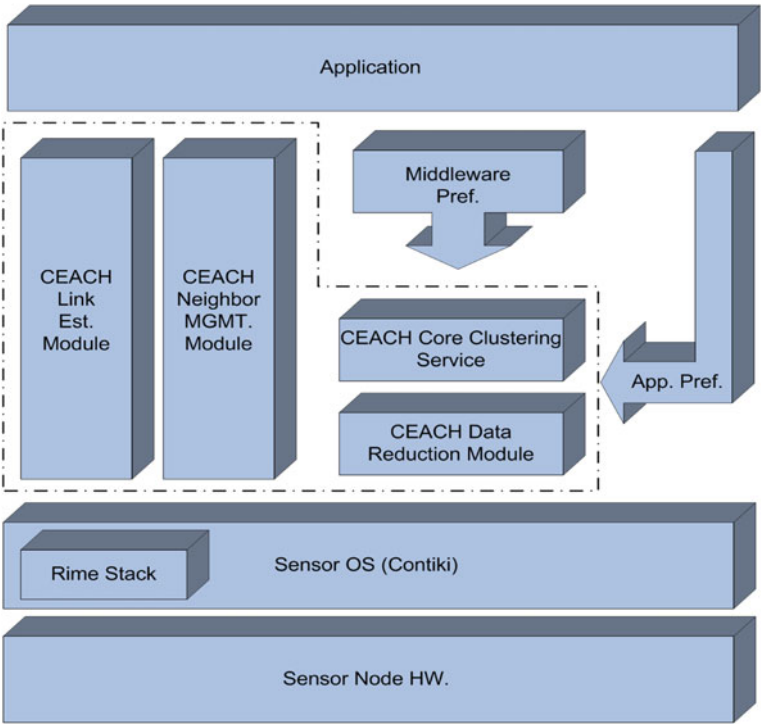


Fig. 6 CEACH middleware architecture for a typical sensor device

hop-distance from the base-station or the sink node are examples of the preferences that an application or a middleware can dictate to CEACH.

The major architectural components that make up CEACH are shown in Fig. 6 and are the following:

- CEACH Link Estimation Module
- CEACH Neighborhood Management Module
- CEACH Core Clustering Service
- CEACH data reduction Module

Before describing details of these components it is appropriate to describe the network model assumed by CEACH that will help put everything in the right context.

1.5.3 The Network Model and Important Definitions

CEACH models a WSN as a directed graph G(V, E) with the set of vertices (V) representing the set of nodes in the network and the set of edges (E) representing communication links between the nodes. The communication links are symmetric

in the sense that if a node (V_1) can communicate with a node (V_2) at a given radio output power level p_r, then the node (V_2) can also communicate with the node (V_1) at the same radio output power level. However, the wireless links, in terms of communication reliability, are not necessarily symmetric which shows their probabilistic nature.

Communication link reliability ($\lambda_{\varepsilon i,j}$) represents the probability of a successful packet transmission from a node (V_i) to a node (V_j) having a communication link $\varepsilon_{i,j}$, where nodes V_i and V_j lie in each others' communication range. Communication reliability on a link is an *asymmetric property* of the link, i.e. it depends on the direction of the link. For instance, if the communication link reliability between two nodes (V_1 and V_2) in the direction $V_1 \rightarrow V_2$ is $\lambda_{\varepsilon i,j}$, it might not be the same in the opposite direction $V_2 \rightarrow V_1$, i.e.

$$\lambda_{\varepsilon_{i,j}} \neq \lambda_{\varepsilon_{j,i}}$$

All edges that are incident upon a vertex (node) represent in-Links of the node from its one hop neighbors. The communication reliability on each of these in-Links is termed as in-Link reliability (in-LR) of the node. Similarly, all edges that emanate from a vertex (node) represent out-Links of the node and communication reliability on these links is termed as out-Link Reliability (out-LR) of the node with its one hop neighbors.

In the multi-hop case, the directed path $\vec{m}$ between two nodes V_i and V_m of length d-hops (3-hops in present case) is represented by $\vec{m}\left(\varepsilon_{i,j}, \varepsilon_{j,k}, \varepsilon_{k,l}, \varepsilon_{l,m}\right)$, where the directed edges from V_i to V_m are given in their order of appearance starting from the node V_i. Since link reliability is a *multiplicative metric*, on a multi-hop directed path, it is the product of the link reliabilities of the constituent links of the multi-hop directed path that constitute the communication reliability of the multi-hop path. Consequently, *End-to-end communication Link Reliability* $ELR(V_i, V_m)^d$ between two nodes V_i and V_m that lie d-hops from each other is the product of the link reliabilities of all the communication links that make up the d-hop path between V_i and V_m.

$$ELR(V_i, V_m)^d = \prod_{\varepsilon_{\vec{m}}}^{d} \lambda_{\varepsilon_{i,j}}$$

The *neighborhood set* of a node (V_i), represented as $N(V_i)$, is composed of the nodes that the node V_i can directly communicate with in a single hop at some specified transmission power level. The neighborhood sets of neighboring nodes overlap with each other. However, two neighborhood sets must have at least one unique member to make them unique, i.e.

$$N(V_i) = N\left(V_j\right), \text{ iff } V_i = V_j.$$

The size of the neighborhood set of a node V_i is called the *degree* of the node $N(V_i)$. The set of all communication links for a node V_i with each of the nodes in

its neighborhood set is represented as $E(V_i)$. The values of the in/out-link reliabilities on each of these links can vary depending upon factors like transmission power, presence of obstacles, multi-path interference, and presence of other devices communicating with each other at frequencies in the ISM band. The average of these values over all the in-links in $E(V_i)$ for a node V_i is termed as the *Mean in-Link Reliability* of the node. Similarly the *Mean out-Link Reliability* of a node is the average of the communication reliability values on all the out-Links of the node.

$$MLR_{in|out}(V_i) = \frac{\sum_{j=1}^{k} \lambda_{\varepsilon_{i,j}}}{k}$$

where k is the size of the neighborhood set of the node V_i.

A *Cluster Head* (*CH*) is a node V_i such that it has the *highest Cluster Head Competence Value* $CHCV_{MLR_{in}(V_i)}$ (defined later) in its neighborhood set. A cluster is composed of any subset (C) of the set of nodes (V), such that all elements of the subset (C) are in either direct communication range of the cluster head (V_i) or are transitively reachable through some member of C, which is called a *Transitive Cluster Head* (*TCH*).

A direct *Cluster Member* (*CM*) is a node V_i such that it can reach the cluster head in a single hop and that its *Cluster Head Competence Value* based upon LR_{out}, $CHCV_{LR_{out}}$, is better than its $CHCV_{LR_{out}}$ with all other cluster heads that are reachable in a single hop from it. If $CHCV_{LR_{out}}$ of a node turns out to be equal for all the cluster heads that are at 1-hop from it, then the node successively compares the components of the $CHCV_{LR_{out}}$ of the *CHs* to choose the best one to join. A *k-level Cluster Member* of a cluster is a node V_i which joins the cluster through some existing direct cluster member or some (k-1)-level cluster member. The k-level cluster members also compare the $CHCV_{LR_{out}}$ of all 1-hop neighbors which are already cluster members or (k-1)-level cluster members and which have made a cluster joining offer.

1.5.4 CEACH Core Clustering Service: Parameter-Space

The core idea of CEACH revolves around structuring the network into node groups called clusters. In order to achieve this goal it uses important attributes related to the individual nodes as well as to the network. It intelligently incorporates different incarnations of these, node and network level attributes, in different phases of its operation. These attributes embody different aspects of the requirements that either come from the application layer or are a result of the evolving network conditions.

Important node and network parameters that can be incorporated in a given routing scheme, either flat or hierarchical, hold the key to its effectiveness. Different routing protocols suggested in the WSN literature mostly focus on reducing

the energy consumption of the nodes and thus strive for prolonging the life of the network. The core idea in the CEACH approach is to focus only on the set of the parameters which is relevant to grouping the nodes in order to structure the network into a clustering hierarchy. In order to better understand the parameters and their significance vis-à-vis the node grouping process, it is appropriate to categorize them either as *node oriented parameters* or *network oriented parameters*.

Node Oriented Parameters

Node oriented parameters, as the name suggests, are the ones which relate with individual nodes in the network. These are the characteristics that either indicate the behavior of a node or represent some of its static property. Here we provide some examples of these parameters along with their significance and usage.

Node Identifier
One of the most fundamental parameters that represent a node is the node identifier. It is a static property of a node that generally doesn't change during its lifetime. Most protocols suggested for WSN and WSAN assume that each node has a unique node identifier which may serve as its address also. However, there are some protocols that disregard it citing the data oriented nature of WSN where collected/sensed data takes precedence over static information about the node which reported this data. The CEACH approach, however, assumes that each node in the network has a unique identifier, which could be an IP address (in case of TCP/IP based Internet of Things setup), or a simple positive integer.

Node Residual Energy
One of the primary concerns in wireless sensor and actuator networks, especially where the nodes are operated on batteries, is to conserve the energy of the nodes with an aim to prolonging the life of the network. Keeping this primary goal in mind, many protocols have built-in mechanisms to make efficient use of this important resource.

The CEACH approach also uses such mechanisms that enable it to ensure balanced energy consumption in the network. This approach ensures that all nodes consume their energy roughly at the same rate by using intelligent load balancing mechanisms. Consequently the chances of developing energy holes in the network are markedly reduced. Additionally, network connectivity can also be ensured for a longer period of time.

Transmission Power
Most sensor nodes, available in the market and which are used in WSN deployments, are embedded with radio modules that can transmit at different output power levels. This node characteristic is exploited by some protocols to conserve battery power also, since radio is one of the biggest consumers of a node's battery. Duty cycling of the radio is an important feature of many MAC protocols [4]. Additionally, many routing schemes also utilize radio duty cycling to ensure

longevity of the battery resource of the nodes. Moreover, a node's neighborhood is determined by its transmission power level. Higher transmission power levels can increase the size of the neighborhood set of a node, since the node can reach and discover more nodes at higher transmission power levels.

The CEACH approach also exploits different transmission power levels available to the nodes. More precisely, the CEACH approach uses different transmission power levels for communication within the clusters (intra-cluster) and across the clusters (inter-cluster). The higher transmission power level is used by the cluster heads to communicate with other cluster heads in the cluster head overlay to transport data in the upstream direction towards the base-station or the sink node. However, if the sink node lies within low-power radio range of a cluster head, then the latter uses low radio output power level to communicate with the former.

Network Oriented Parameters

Node Degree

Node degree, simply put, is the number of nodes that a given node can communicate with at a specific radio output power level. Since the neighborhood of a node depends on the radio transmission power level used for communication and since CEACH uses two different power levels for intra-cluster and inter-cluster communication, this results in two different neighborhoods for a node depending upon its role in the communication hierarchy. For ordinary nodes, i.e. cluster members, there is a low power neighborhood, where low power is a defined power level known to each node. For nodes in the cluster head role, there are two neighborhood sets and consequently two values for Node Degree. Like the nodes in the ordinary cluster member status, the cluster heads have a low power neighborhood. However, for inter-cluster communication, they have a high power neighborhood also that consists of those cluster heads which they can communicate at the defined high power level.

Network Size

Total number of nodes in the network defines the network size. Some centralized clustering approaches require the network size to be known beforehand for the clustering protocol to work. However, the distributed clustering approaches like the one used by CEACH also, because of their very nature, don't require network size to be known beforehand.

Hop Distance

One important parameter for any multi-hop protocol is the number of hops that a given message needs to travel before arriving at its final destination. It also indicates the volume of network traffic as higher number of hops means higher number of message transmissions. Many routing protocols try to reduce the hop distance between a given node and the base station while forming a routing tree. In a multi-hop clustering protocol, hop distance is of significance in defining not only the intra-cluster topology but also the inter-cluster topology. A communication

path with lesser number of hops should be preferred over a path with higher number of hops provided the average hop-length in the shorter path is not longer than the average hop length in the longer path [8].

Sensor Data Correlation
Since a clustering hierarchy can be exploited for data aggregation and date fusion purposes, clusters with the nodes which observe more cohesive sensor data can perform better data fusion jobs than the ones where there is little correlation between sensor readings of the nodes. Some recent approaches that emphasize data fusion using clustering structures do emphasize the use of spatial data correlation as one of the parameters while forming clusters. On the other hand, it could be argued that parameters like temperature and humidity don't change that abruptly over shorter distances except for in controlled environments; therefore, sensor data is correlated anyway in a given neighborhood in the network.

Link Reliability
Generally speaking, incorporation of link reliability information by routing and clustering protocols can bring dividends. Since it is the quality of a communication link that really defines communication cost over a link than any other factor, it underscores the importance of considering link reliabilities in the clustering process. Traditionally, metrics like hop-count, round-trip time and latency have been used to estimate the quality of a routing path. These traditional metrics, however, have shown to be of lesser significance in estimating quality of wireless links in WSN [23].

Specifically, the minimum hop-count metric has been used extensively in wireless networks to choose better routing paths in several protocols. However, Expected Number of Transmission count (ETX) metric suggested by [8] have been shown to be better than the simple hop-count metric. There are quite a few metrics that have been suggested in the WSN reliability literature that could be used as such for this purpose. They have been included here along with their significance in the clustering process and their ease or difficulty of assessment. Important aspects of a metric are: the ease of its measurement, its effectiveness and its representativeness of the quantity that it measures. A given metric which is straightforward to determine but is not representative of the physical quantity that it is supposed to measure is not a good metric. With the same principle in mind, we list the available link reliability metrics along with their significance, ease of measurement and representativeness (see Fig. 7).

Hardware Oriented Link Quality Estimators
One category of link quality estimators are called the hardware based estimators, since their estimation is based upon the data provided by the radio hardware. They are very conveniently and rather easily obtained from the radio chip on the node. They, however, have been criticized for not being truly representative of the link quality [23]. Some of the more well known hardware oriented link assessment metrics are:

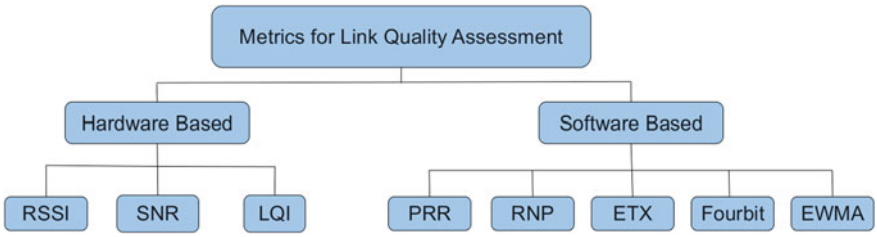


Fig. 7 Metric space for link reliability assessments in ad-hoc wireless networks

- Received Signal Strength Indicator (RSSI)
- Signal to Noise Ration (SNR)
- Link Quality Indicator (LQI).

Software Oriented Link Quality Estimators
Software oriented link quality metrics are shown to be more representative in describing the quality of a link. These metrics are generally harder to obtain than their hardware counter-parts. But their suitability to representing quality of a link more accurately outweighs the difficulty that one faces in assessing them.

- Packet Reception Ratio (PRR)
- Required Number of Packet Transmissions and Retransmissions (RNP)
- Expected Number of Transmissions (ETX)
- Fourbit (4Bit)
- Exponentially Weighted Moving Average (EWMA)
- Window Mean with Exponentially Weighted Moving Average (WMEWMA) (t, α).

Packet Reception Ratio (PRR)
Simply stating, Packet Reception Ratio (PRR) on a given link is the ratio between the number of packets sent over the link to the number of packets that is actually received. Since PRR doesn't take into account the underlying distribution of the packet losses over a given link, it tends to overestimate the link quality.

Required Number of Packet Transmissions and Retransmissions (RNP)
RNP is a metric suggested by [7] for assessing quality of low power wireless links. According to the authors, RNP shows good co-relation with the temporal properties of a wireless link. It measures the total number of transmissions needed to deliver a packet over a wireless link in a network where Automatic Repeat Request (ARP) for an undelivered packet is enabled. One issue with this metric is that it assumes that the underlying packet distribution for the wireless links in the network is known apriori.

Expected Number of Transmissions (ETX)
De Couto et al. [8] suggested Expected transmission count (ETX) on a given wireless link as an estimate of the number of packet transmissions and retransmissions, including acknowledgements, that are expected between two nodes for a

successful unicast packet transmission. It is relatively a simple metric to compute and precisely for this reason is quite popular link estimation metric in the research community. ETX for a link, between two nodes A and B, can be computed as follows:

$$ETX = 1/PRR_{A\rightarrow B} \times PRR_{B\rightarrow A}$$

It is a beacon based metric, in the sense that a node uses beacons, and not data packets, to estimate link quality using ETX. Packet Reception Ratio in each direction between two nodes need to be computed in order to know ETX for that link.

Fourbit (4Bit)
It is a link estimation metric that utilizes information from PHY, Link, and Network layers. A link assessment metric should be accurate/precise, responsive and should provide useful information. Its main argument is that PHY layer provides per packet information and misses the temporal variations in the channel quality. The PHY layer provides the channel quality information for the successfully received packets, thus over estimating the quality of a link. At the Link layer the acknowledgement mechanism can be used to keep a count of the acknowledged and unacknowledged packets thus generating a very accurate estimate of the quality of the link. *ETX* operates at this level, but it uses dedicated probe broadcasts to assess the link quality assuming that these probe broadcasts show the same behavior or follow the same pattern as the actual data traffic. Moreover, the network layer has the routing information, thus it has the knowledge which links are most important to routing.

However, estimating link quality at the network layer is inefficient and slow to adapt to the changes in the link quality.

Window Mean with Exponentially Weighted Moving Average (WMEWMA)(t, α)
This metric measures link quality as a percentage of the packets that are received undamaged over a given wireless link. It tries to compute an average success packet delivery success rate during a given time window of duration (t). *WMEWMA* then tries to smoothen this packet delivery success rate with an Exponentially Weighted Moving Average (*EWMA*). The tuning parameters used by *WMEAMA* are the time window for the current delivery success rate measurements, t, and a history contribution factor, α. Since this metric incorporates history into its measurements, it is considered to represent quality of a given link better than the metrics that don't incorporate history factor. The simplicity of the metric along with less memory requirements and incorporation of the history factor are considered its advantages. However, the metric has received some criticism for requiring at least W packet transmissions before making an estimate of the link quality.

1.5.5 CEACH Link Estimation Module (CLEM)

Different link estimation metrics described in the previous section highlight the significance that is associated with link assessments in routing and topology maintenance in wireless ad-hoc networks. The incorporation of link quality assessments while making routing or topology maintenance decisions has been shown to pay dividends in terms of higher message delivery ratios [12, 29] and longer network life times. However, some clustering protocols, like the one suggested by [36], try to locate centers of mass in the network which act as cluster heads. These centers of mass are the nodes that lie in the high node density regions and at a central location in these high density regions where they are supposedly more reachable from other nodes.

Additionally, several routing protocols are based upon the assumption that energy consumption in a wireless network is a function of the distance between communicating nodes. For precisely this reason many protocols use hop-count or Euclidean distance between nodes to optimize parent selection in the network. However, these approaches are naïve in assuming that the energy consumption in the network is a function of the distance between communicating nodes. They ignore the fact that such centers of mass in the network could not only create hot-spots, having high message traffic and high energy consumption, but can potentially be also low average link reliability regions.

CEACH approach shifts markedly from the centers-of-mass approach and presents the idea of centers of high link reliability regions. These are the regions in the network where the mean link reliability is high. Figure 8 depicts the process of locating centers-of-high-link-reliability regions in the network.

Additionally, the nodes' coloring scheme indicates their residual energy, another parameter used in CEACH for determining cluster heads as well as communication partners for upstream data traffic.

The link estimation module in CEACH is responsible for locating centers-of-high-link-reliability. This module, in principle, can use any of the link estimation metrics presented in the previous section. That said, for the results presented later in the chapter, it uses *WMEWMA* to estimate link quality.

1.5.6 CEACH Neighborhood Management Module (CNMM)

Neighbor discovery in wireless ad-hoc networks like wireless sensor and actuator networks is not a trivial task. Even if it is assumed that all the nodes are static, still the nodes can die by running out of their battery power. The wireless links are also prone to changes in their quality of service and there are occasions when a wireless link is considered either too expensive to communicate over due to bad link quality or completely unavailable due to outage of communication over it. These factors make neighbor discovery a non-trivial problem in WSN and WSAN. Therefore,

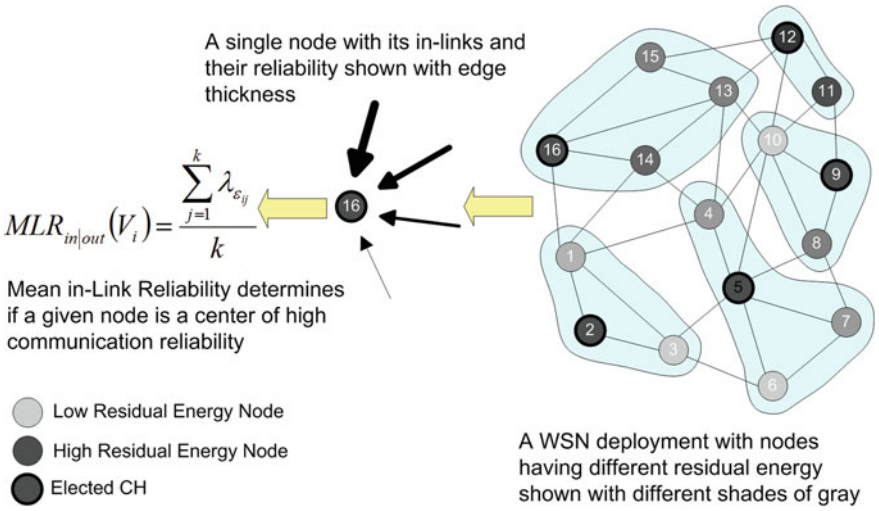


Fig. 8 CEACH approach—locating the centers of high communication reliability in the network

the neighbor-lists need to be kept up-to-date using some agile mechanism. Using explicit beacons at fixed intervals add to the communication overhead especially if the beacon interval is too low. Having a long beacon interval, on the other hand, make the neighbor table entries unreliable. *CNMM* uses Neighbor Discovery Messages (*NDM*) for the purpose. These messages are multi-purpose, as they not only act as a neighbor discovery mechanism but also as a means to share other information like residual energy of the node within its neighborhood. This scheme of using a type of a message to achieve multiple goals keeps the overhead low.

CEACH Data Reduction Module (CDRM)

Since in-network processing is crucial to conserving battery power of the nodes and in turn improving the network lifetime, CEACH provides in-network processing in the form of data reduction functions. Data reduction in this context refers to removing redundancy from data, e.g. threshold based data reporting, as well as aggregation functions like *MIN*, *MAX*, and *AVERAGE* etc. Since these functions can be applied on a dataset, applications can dictate a time window called an *EPOCH* which is used by the *CDRM* to apply these functions. Moreover, CEACH offers application of data reduction functions on a spatio-temporal basis, thereby meaning that these functions can be applied on the level of a node as well as on the level of cluster heads. The provision of these functions at the level of cluster heads provides for the spatial data reduction in the network. The combination of spatial and temporal data reduction functions make CEACH middleware appropriate for many sensor network applications where these features are required.

1.5.7 CEACH Clustering Service (CCS)

The core CEACH Clustering Service (*CCS*) provides two interfaces and also uses two interfaces. The interfaces that it provides are the following:

- Interface to the application running over it and using its services, so that the application can give its preferences over to the *CCS*.
- Interface to any middleware service that is using its services, so that the middleware can give its preferences over to the *CCS*.

The interfaces that *CCS* needs are the following:

- Interface from the Neighbor Management Module for gaining access to the up-to-date neighbor list.
- Interface from the Link Estimation Module to gain access to the up-to-date link quality statistics over various links that the node has within its neighborhood.

CCS uses a cost metric to evaluate the suitability of a node for becoming a cluster head. The metric is named Cluster Head Competence Value (*CHCV*) and is explained in the next section.

CEACH Clustering Service: The Cost Metric

CCS uses a weighted composite metric (*CHCV*) that incorporates important node and network oriented attributes. The constituent attributes of this metric are converted to indices, whose values vary between 0 and 1, using schemes that ensure the desired contribution and impact of the attributes in the metric.

Residual Energy Index (REI)
If a WSN is structured into clusters, the nodes that form a cluster generally report their data to the elected *CHs*, which normally applies spatial aggregation function(s), the type of aggregation function applied depends on the application preferences, and forward these data to the base station either directly in a single-hop or multiple hops. The role of the cluster head, thus, has some additional responsibilities which put higher demands on their already constrained resources. In some approaches, *CHs* are assumed to have better resources, e.g. long-life batteries, than the normal nodes in the network. However, most clustering approaches consider all nodes to have homogeneous resources. Therefore, it is wise to consider the residual energy of the nodes while choosing *CHs*. The approaches that consider residual energy of a node while choosing cluster heads, however, mostly try to relate power consumption with the distance between the nodes and not with communication link reliability.

CCS also attaches great importance to the Residual Energy (E_{RE}) available to the nodes and uses the information about it in different phases of its operation. Many approaches that consider residual energy in structuring the network, assume that all nodes have the same initial energy. However, unlike these approaches,

CCS doesn't make any prior assumptions about the energy homogeneity of the nodes. It also takes a broader view of power consumption and doesn't confine it to just the distance between the communicating nodes. This point is worth noting, since most approaches relate the power consumption to the distance between the communicating nodes and ignore the quality of the link between them altogether. *CCS* also defines a threshold value for the energy ($E_{TE)}$ of the nodes which marks the bare minimum energy level a node should possess, if it is to take part in the cluster head election process. This is a design parameter of the protocol and can be tuned appropriately to suit the application needs and expected load on the *CHs*. As soon as the E_{RE} of a cluster head falls below the E_{TE}, it gives up its cluster head status and calls for re-clustering the network.

In each round of clustering, the E_{TE} value is reduced by a specific percent of its value in the last round. This is necessary to ensure that there are nodes that have higher E_{RE} than the E_{TE}, which enable them to contest the cluster head election. It could, however, be argued that a cluster head whose E_{RE} falls below the E_{TE}, might become a cluster head again in the subsequent round, if E_{TE} is dropped by a certain amount in each round of clustering.

Algorithm 1 DETERMINE REI

1: **if** ($E_R > E_T$) **then**
2: REI = ((E_R - E_T)/ E_R)
3: **else**
4: REI = 0.001

In reality it could seldom happen, since *CHs* consume more energy in carrying out additional duties. So, even if a cluster head whose E_{RE} has fallen below the E_{TE} and which has called for re-clustering the network, subsequently has a higher E_{RE} than the E_{TE} after lowering the later, it might not necessarily be the best node in its neighborhood to assume the cluster head role again. Each node is assigned an REI between 0 and 1, as is given in the Algorithm 1, where 1 is the best value. However, if a node's E_{RE} is below the E_{TE}, it is assigned a minimum value of 0.001 to eliminate its chances of becoming a cluster head.

Node Degree Index (NDI)

Another aspect that the *CCS* incorporates in the *CHCV* metric is the size of the node's neighborhood, i.e. *Node Degree*. It gives an indication of the possible size of the 1-hop membership of a cluster, should the node become a cluster head. Therefore, in the cluster head election and cluster joining phases of the clustering process, node degree is used in the *CHCV* metric. *CCS* uses the notion of *Ideal DEGree* (*IDEG*), which serves the purpose of giving more importance to the nodes that have the desired degree in becoming *CHs*. This could be used as a load balancing mechanism in the cluster joining phase. Similar to *REI*, each node

computes Node Degree Index (*NDI*) (as depicted in Algorithm 2) whose value varies between 0 and 1, with 1 being the best value. This scheme ensures that the nodes with degrees equal to the *IDEG* get a maximum value of 1 and other nodes always get a value lower than 1. In the inter-cluster communication phase, however, *CCS* doesn't use node degree to compute *NDI*.

Algorithm 2 DETERMINE NDI°

```
1: if node degree = IDEG then
2:         NDI = 1.0
3: else if node degree > IDEG then
4:         NDI = IDEG/node degree
5: else
6:         NDI = node degree/IDEG
```

There *CCS* uses each cluster head's actual 1-hop membership count instead. This scheme could be exploited to the benefit by *CCS* in two different ways. Firstly, if *CHs* strive for a higher degree of aggregation in the network, they can choose those *CHs*, as their upstream neighbors, which have a higher 1-hop membership count. On the other hand, if the goal is to balance the load and consequently energy consumption, then *CHs* with a lower 1-hop membership count could be chosen as upstream neighbors because they have less cluster management load and would possibly have a higher E_{RE} too.

Link Reliability Variants

The *CHCV* metric slightly differs in each phase of the *CCS's* clustering process. In the cluster head election phase, *Mean in-Link Reliability* (MLR_{in}) is used to calculate it and is accordingly represented as $CHCV_{MLR_{in}(V_i)}$.

$$CHCV_{MLR_{in}(V_i)} = REI \times IF_{REI} + NDI \times IF_{NDI} + MLR_{in}(V_i) \times IF_{MLR} \quad (1)$$

The inclusion of MLR_{in} in the above Eq. (1) guarantees that the nodes which are strongly connected in their 1-hop neighborhood have higher chances of getting elected as *CHs*. Please note that *CCS* includes the in-link reliabilities to compute *MLR* because it assumes that most of the information flow, in a clustered network, is in the upstream direction, i.e. from nodes to the *CHs* and then towards the sink. It, however, doesn't limit the application of *CCS* protocol, in any way, to the scenarios where information flows in both directions. In such cases, *MLR* is computed by considering both in and out link reliabilities of the edges that are incident upon or are emanating from a node.

One important question arises because of the inclusion of MLR_{in} or an average *ETX* value of all the links that a given node has in its 1-hop communication graph. Since the successful election of a node as a cluster head doesn't necessarily mean that all the nodes in its 1-hop communication graph will join it to form a cluster, one can raise questions on the significance of using *MLR* as an important

parameter in electing *CHs*. If a node's contribution in *MLR* to any of its neighbor's increases the latter and then in the cluster formation phase this node doesn't join the resulting cluster and instead joins some other neighboring cluster, even then it doesn't affect the validity of the cluster head election. In the cluster formation phase, each node, lying at 1-hop from the elected *CHs*, uses (Eq. 2) which is obtained by replacing MLR_{in} in Eq. (1) with out-Link Reliability of the node with the elected *CHs* to evaluate their cluster joining offers.

$$CHCV_{LR_{out}(V_i)} = REI \times IF_{REI} + NDI \times IF_{NDI} + LR_{out}(V_i) \times IF_{LR_{out}} \quad (2)$$

The nodes that are not in direct communication range of any cluster head join clusters transitively through existing *CMs* or *TCMs*. Their cluster joining decision is also based upon a mix of attributes which is represented by Eq. (3). In the same way, when the inter-cluster communication paths are formed, *CHs* use $CHCV_{ELR_{in|out}}$, which is obtained by replacing $\mathrm{MLR_{in}}$ in Equation with end-to-end out/in-Link Reliability of the cluster head with the sink node, to choose the best upstream neighbor.

$$CHCV_{ELR_{out}(V_i)} = REI \times IF_{REI} + NDI \times IF_{NDI} + ELR_{out}(V_i) \times IF_{ELR_{out}} + \frac{1}{\zeta} \times IF_{\zeta} \quad (3)$$

It is evident from the above three Eqs. (1, 2, and 3) that *CCS* incorporates relevant notions of link reliabilities in all phases of the clustering process. Additionally, it incorporates node's residual energy to make sure the nodes with higher energy get preference in all phases of the *CCS* clustering process. The impact of the constituent parameters of *CHCV* can be controlled by varying the values of the *Impact Factors* (*IFs*) in the above equation. The value of each of these *Impact Factors* varies between 0 and 1, and like any weighted average their sum equals 1. *CCS*, however, avoids assigning a value of 0, which would effectively remove the influence of that particular parameter from the metric. However, if a given application demands maximum focus on only one of these parameters, then an *Impact Factor* of 1 could be assigned to that parameter. The incorporation of *Impact Factors* in *CHCV* adds flexibility and allows one to choose an appropriate mix which suits the given application scenario.

1.5.8 Link Reliability and Hop-Distance

The third component of the *CHCV* metric depends upon a node's strength of communication links with its 1-hop neighbors, i.e. link reliability. Two types of metrics could be used to assess reliability of a communication link, namely, hardware based and software based, as described previously. Examples of the hardware based metrics are Link Quality Indicator (*LQI*), Received Signal

Strength Indicator (*RSSI*), and Signal to Noise Ratio (*SNR*). These metrics are easy to obtain directly from the radio hardware. However, they are calculated by the radio hardware only for the successfully received packets and that too on the basis of the first 8 symbols of the received packet. They have also been shown to be inadequate to properly represent the quality of a communication link. Among the software based link reliability metrics, the most well known are the Packet Reception Ratio (*PRR*) and its derivatives. *CCS* uses *WMEWMA* to quantitatively represent the reliability of a communication link. The hop-distance, (represented by ς in Eq. 3) between either a k-level *CM* and its cluster head or a node and its upstream neighbor in the inter-cluster communication, can influence communication costs. Although the multiplicative nature of the end-to-end link reliability takes care of the hop-distance implicitly, it fails to distinguish between two multi-hop routes where link reliability has a maximum value of 1. *CCS*, therefore, incorporates hop-distance into the *CHCV* metric (Eq. 3) wherever multi-hop paths are involved. This enables the nodes, while evaluating either k-level cluster membership offers or potential upstream neighbors in the inter-cluster communication, to minimize their hop-distance and, thus, lower the communication cost.

1.5.9 CEACH Clustering Service: The Cluster Head Election

At the beginning of the *CCS* cluster head election protocol each node is in the *UC* state. The first task is to determine the neighborhood set $N(V_i)$ of each node as well as the initial link reliabilities within the neighborhood set. It is achieved by randomly broadcasting Neighbor Discovery Messages (*NDMs*) with one of the lower transmission power levels P_{T_X} available to the node. The higher transmission power levels are allocated for inter-cluster communication. During this initial phase of determining its neighborhood set, each node sends "n" such broadcasts where "n" is a positive number whose value could be chosen depending upon the degree of certainty required in determining the link reliability.

These repeated random broadcasts serve two purposes. For one, they help in determining accurate $N(V_i)$ of a node. Secondly, they help determine the link reliability of each node with the members of its neighborhood set using *WMEWMA* metric. In the first phase of clustering, i.e. cluster head election, each node determines its Mean in-Link Reliability (MLR_{in}) which is based upon the initially exchanged *NDMs*.

However, upon re-clustering the network, the nodes utilize the messages exchanged during the normal operation of the network to compute both in and out-link reliabilities again using the *WMEWMA* metric. This takes into account the time varying nature of the link reliability values of the links, since the information used to compute them is gathered over a time window of duration (t). This also adds to the adaptive nature of CEACH, since between two clustering periods each

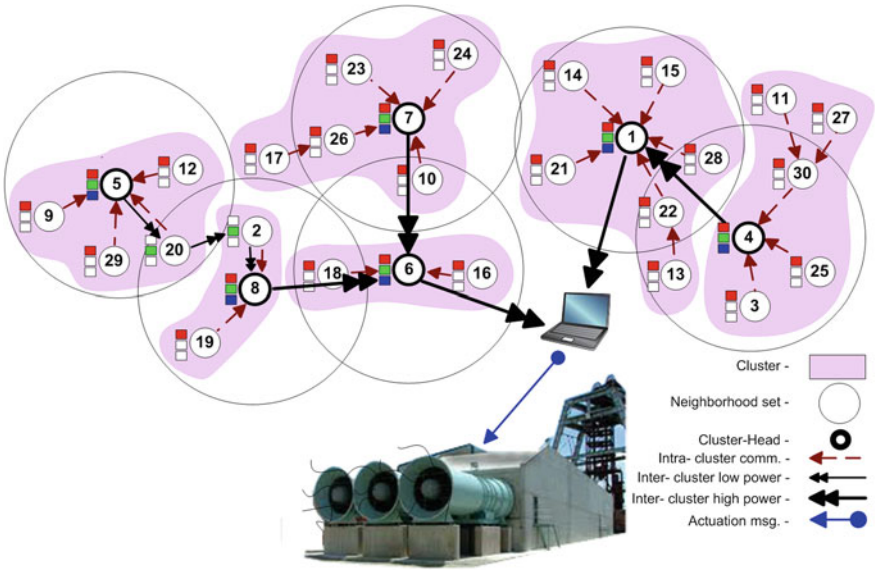


Fig. 9 A WSAN structured by CEACH middleware service, showing a possible actuation scenario of industrial scale ventilators to implement air quality control mechanisms in hazardous work environments

node collects statistics on its successful or otherwise message transmissions with its 1-hop neighbors. These statistics are shared within the 1-hop neighborhood in each new round of clustering, so that they could be used to recompute link reliabilities. Thus each node has more reliable data on its communication links which enables it to make better decisions in each successive round of clustering. One of the design parameters of CEACH, which has been included to ensure that only high energy nodes compete for becoming *CHs*, is the threshold energy (E_{TE}). Figure 9 shows a WSAN which has been structured by CEACH middleware service as well as a possible actuation scenario in an industrial setting.

Algorithm 3 CLUSTERING (PHASE-1: CH ELECTION)

1: Variables: *NDM* = Neighbor Discovery Message, *ICH* = I Am CH Msg., *SNCH* = Single Node *CH*
2: randomly broadcast *NDM* '*n*' times
3: wait, for specific time, to hear *NDM* announcements
4: **if** an *NDM* announcement is received **then**
5: Request *CNMM* to add sender to the neighbor-list
Request *CLEM* to start assessing *LR* on the link
6: **if** my neighborhood set = ϕ **then**
7: Set my_status = *SNCH*
8: **else**
9: **if** it is a re-clustering call **then**
10: Get *LR* on each link from *CLEM* & share it with all neighbors
11: Get *in-LR* on each link from *CLEM*
Compute MLR_{in}
12: **if** my $E_{RE} < E_{TE}$ **then**
13: Set my_status = *UC*
14: **else**
15: Set my_status = *CHC*
16: Share MLR_{in}, E_{RE}, D_{V_i} within 1-hop Neighborhood
17: Compute your own and each neighbor's $CHCV_{MLR_{in}}$
18: **if** my $CHCV_{MLR_{in}}$ > all neighbor's $CHCV_{MLR_{in}}$ ^ my_status = *CHC* **then**
19: Set my_status = *CH* and broadcast *ICH* Msg.
20: **else if** my $CHCV_{MLR_{in}}$ = any neighbor's $CHCV_{MLR_{in}}$ ^ my_status = *CHC* **then**
21: Compare your MLR_{in}, E_{RE}, D_{V_i}, *NID* in the same order with that of your 1-hop neighbors
22: if you are best, set my_status = *CH* and broadcast *ICH* Msg.
23: **else**
24: set my_status = *UC* and wait for *ICH* Msg.

This is the bare minimum residual energy of a node which allows it to assume the role of a cluster head. Since *CCS* wants to achieve well balanced clusters that don't vary in size greatly and which can also minimize the interference (should the application desires some *TDMA* based *MAC* scheme inside the clusters), *CCS* can achieve that by assigning an appropriate value to the *IDEG* design parameter. This parameter defines *CCS's* preferred cluster size. In the experiments, it has been assigned a value of 4 which simply means that the nodes having a degree of 4 are favored in becoming cluster heads. This value could be changed easily, if the

application requirements are to have clusters of some specific size. The way *CCS* utilizes it in the cluster head election phase is shown in the Algorithm 3.

It ensures that a node, having a degree closer to *IDEG*, is preferred to assume the role of cluster head in its neighborhood. The three parameters that make up the metric, which is used in cluster head election phase of the clustering process, are shared with the nodes in the 1-hop neighborhood. Each node in the *CHC* state determines if it is the best suited node to assume the role of a cluster head by comparing its $CHCV_{MLR_{in}(V_i)}$ value with that of its neighbors. Ties are broken by comparing the constituent parameters of the $CHCV_{MLR_{in}(V_i)}$ metric in the desired order. Currently, *CCS* resolves ties by comparing the nodes' MLR_{in}, E_{RE}, and *NID* respectively. This order is based on the observation that a node having relatively higher energy can dissipate it quickly, if it has poor link reliability with its neighbors in its 1-hop neighborhood. Finally, the best node assumes the cluster head role and broadcasts its cluster joining offer to its 1-hop neighbors to form clusters. Figure 10 shows the *CCS's* clustering phase in a state-transition diagram.

1.5.10 Cluster Formation

In the cluster formation phase of *CCS*, the nodes wait, a specific amount of time, to hear cluster head announcements (*cluster joining offers*). Upon hearing these offers, each node selects the best offer using the $CHCV_{LR_{out}(V_i)}$ metric. In this metric out-Link Reliability is used, since it is more relevant in a clustering hierarchy as most of the communication takes place from the nodes to their respective cluster head. It, however, can be replaced with a metric like *Estimated Transmission count* (*ETX*) (or *WMEWMA*) that includes both in and out link reliabilities, should the application requirements necessitate so.

If two offers have the same value for $CHCV_{LR_{out}(V_i)}$, nodes use *out-link reliability*, *degree*, E_{RE}, *NID* respectively to break the tie. The nodes that lie outside 1-hop range of the elected *CHs* cannot hear any cluster head announcements directly.

They, however, receive offers from existing *CMs* or *TCMs* to join a cluster transitively. Once a node hears such offers, it uses $CHCV_{LR_{out}(V_i)}$ to evaluate them. Here the notion of end-to-end link reliability of the complete d-hop path to the cluster head is used instead of a greedy approach whereby out-link reliability to the nodes offering transitive cluster membership is used. *CCS* chose this end-to-end approach because of its obvious advantage over the greedy approach. The greedy approach would suffer if the multi-hop communication link up to the cluster head has low link reliability after the immediate neighbor. In case of a tie between two offers, nodes use ELR_{out} between themselves and their cluster head, *hop-distance* to the cluster head, *degree*, E_{RE}, and *NID* to break the tie. The Fig. 11 shows in/out-link reliabilities on the edges between nodes. The *ELRs* are shown with lines that span more than one edge length. Node-12 is faced with a complex cluster joining decision. It has no CH in its 1-hop communication range. After receiving

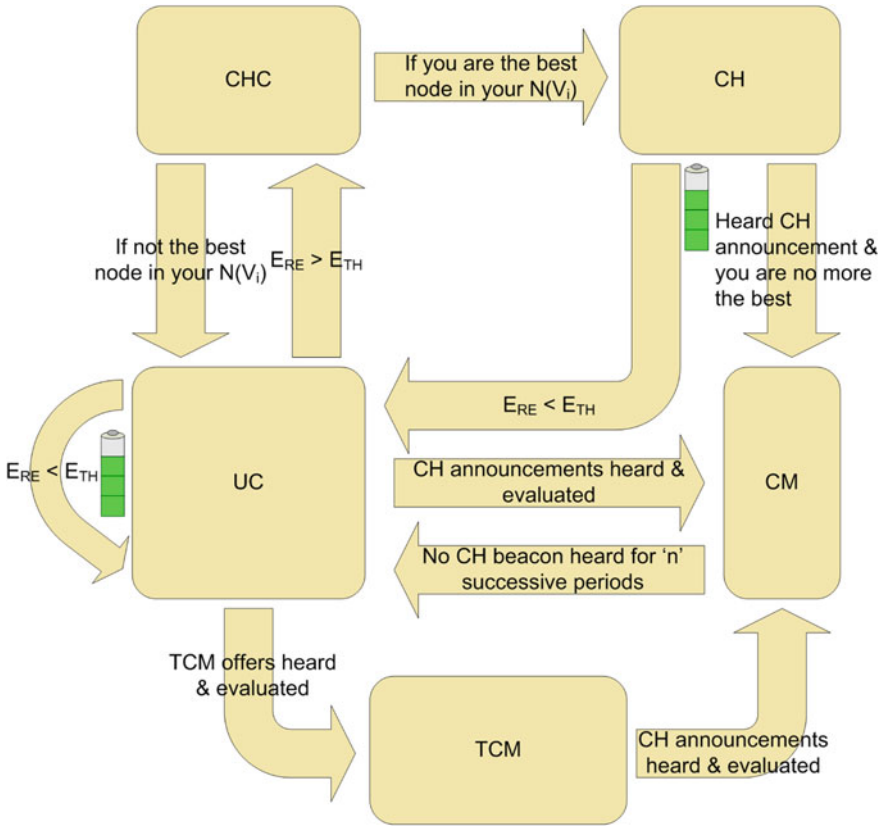


Fig. 10 State-transition diagram depicting different states a node goes through during the execution of *CCS*

transitive cluster joining offers, it finds out that there is a tie on the basis of the value of the $CHCV_{ELR_{out}(V_i)}$ metric in all the three offers that it receives. If it were to take a greedy approach, it should accept that offer with the highest *out-LR*, which in the present case is not applicable, since on all three out-links it has the same *out-LR* (0.5). On the contrary, it takes an end-to-end approach and compares the offers on *ELR* basis. The *ELRs* on two paths have the same value of 0.35. In order to break this tie, it successively compares the other parameters that make up the $CHCV_{ELR_{out}(V_i)}$ metric, ultimately breaking the tie using E_{RE}. Therefore, the end result is that the node-12 chooses *CH*-5 via *TCH*-4 instead of *CH*-7 via node-3, since the node-4 has a higher E_{RE} than the node-3, assuming everything else being equal.

Algorithm 4 CLUSTERING (PHASE-2: CLUSTER FORMATION)

1: Variables: *CJN* = Cluster Joining Notification (with the chosen *CH id*), *SCJ* = Searching Cluster to Join Msg., *TCMO* = Transitive Cluster Membership Offer (unicast), *SN* = sub_neighbor, *TCJN* = Transitive Cluster Joining Notification (with the chosen *TCH id*)
2: **if** *ICH* Msg. is received **then**
3: **if** my_status != *CH* **then**
4: wait, for specific time, to receive more *ICH* announcements
5: Evaluate cluster membership offers using $CHCV_{LR_{out}}$, *LRO*, D_{V_i}, E_{RE}, *NID*, in the same order
6: Accept membership offer
7: Set my_status to *CM* and broadcast *CJN*
8: **else if** my_status = *CH* **then**
resolve conflict using *CH* Election criteria in Algorithm-3
9: **if** my_status = *UC* • no *ICH* Msg. is received **then**
10: Broadcast *SCJ*
11: **if** an *SCJ* Msg. is received **then**
12: **if** my_status = *CH* **then**
13: unicast *ICH*
14: **else if** my_status = *CM* • *TCM* **then**
15: unicast *TCMO*
16: **if** an *TCMO* Msg. is received and my_status = *UC* **then**
17: Wait, for specific time, to receive more *TCMOs*
18: Evaluate *TCMOs* using $CHCV_{ELR_{out}}$, ELR_{out}, • , (D_{V_i}), E_{RE}, *NID*, in the same order
19: Accept best *TCMO*
Set my_status to *TCM*
Broadcast *TCJN*
20: **if** an *CJN* Msg. is received **then**
21: **if** my_status = *CH* **then**
22: **if** *CJN* sender's *CH* are you **then**
23: Request *CNMM* to mark the sender as your *CM*
24: **else if** sender is already your *CM* and *CJN* sender's *CH* are not you **then**
25: Request *CNMM* to unmark the sender as your *CM*
26: **else if** (my_status = *CM* • *TCM*) • sender = my *SN* **then**
27: Request *CNMM* to unmark the sender as your SN
28: **else if** (my_status = *CM* • *TCM*) • the sender is my *CH* **then**
29: set my_status = *UC* and broadcast *SCJ*
30: **if** an *TCJN* Msg. is received **then**
31: **if** my_status = *CM* • *TCM* **then**
32: **if** *TCJN* sender's *CH* is me **then**
33: Request *CNMM* to mark the sender as *SN*
34: **else if** *TCJN* sender's *CH* is not me • sender is my *SN* **then**
35: Request *CNMM* to unmark the sender as *SN*

The usefulness of *TCHs* can be exploited by *CCS* if it makes them share the cluster management load with their *CHs* by registering the *TCMs* with themselves and bearing all the responsibility of aggregating and forwarding their data to the *CHs*. In this way, the *CHs* receive one aggregated message from a *TCH* which represents data of both TCH and its sub-neighbor(s) or the nodes which are *TCMs* via it. In case a *CH* loses its status, the *TCMs* can be informed of this change by the *CHs*. The adaptive nature of the *CCS* allows nodes to switch clusters, if they receive a better offer, even after they have accepted an earlier cluster joining offer. This, however, raises some concern regarding the network state consistency, as multiple *CHs* or *TCHs* can have a node listed as their *CM* or *TCM* respectively. This issue is resolved in *CCS* by exploiting the broadcast nature of the wireless communication as explained below.

Algorithm 5 CLUSTERING (PHASE-3: INTER-CLUSTER COMMUNICATION)

1: Variables: *SCHICC* = Searching *CH* for Inter-Cluster Communication at High Power (*HP*), *FCHICC* = Forwarded *SCHICC* Msg. at *HP*, *BACKFSHP* = *Block ACK* From Sink at *HP* (with Msg. count of *FCHICC* that sink overheard (*IACK*)), *USN* = Upstream Neighbor, *SLPNICC* = Searching Low Power Neighbor(*LPN*) for Inter-Cluster Communication, *FLPNICC* = Forwarded *SLPNICC* Msg. at *LP*, *BACKFSLP* = *Block ACK* From Sink at *LP*(with Msg. count of *FLPNICC* that ink overheard (*IACK*)), *USN* = Upstream Neighbor, *HTS* = hops to sink
2: **if** my_status = *CH* **then**
3: **if** an *SCHICC* is received **then**
4: Request *CLEM* to start assessing *LR* on the link
Mark the sender (Sink) as your *HP USN*
Set *HTS* = 1,
Wait, for specific time, to receive more *SCHICC* Msgs. from Sink
5: Request *CLEM* to compute *in-LR* with sink
Broadcast *FCHICC* n times
6: **if** an *BACKFSHP* Msg. is received **then**
7: Request *CLEM* to compute *out-LR* with the sink
8: **if** an *FCHICC* is received • sender's *USN* are not you **then**
9: Request *CLEM* to start assessing *LR* on the link
Store the sender's *HTS*
Wait, for specific time, to receive more *FCHICC* Msgs.
10: Get *in-LR* with the sender from *CLEM*
11: Evaluate *IC Comm*. offers using $CHCV_{ELR_{out|in}}$, $ELR_{out|in}$, •, E_{RE}, *NID*
12: Mark selected *CH* as your *HP DSN*, set *HTS* = selected *CH's HTS* + 1
13: Broadcast *FCHICC* further '*n*' times
14: **else if** an *FCHICC* Msg. is received • sender's *USN* are not you **then**
15: Request *CNMM* to mark the sender as your *USN*
Request *CLEM* to assess *in-LR* with the sender

16: **if** an *SLPNICC* is received **then**
17: **if** the sender is the sink node **then**
18: Request *CLEM* to start assessing *LR* with the sink
Request *CNMM* to mark the sink as your *LP USN*
Set *HTS* = 1
Send *Ack*. to the sink
19: Wait, for a specific time, to hear more *SLPNICC* from the sink
20: Get *in-LR* with the sink from *CLEM*
21: **if** an *BACKFSLP* Msg. is received **then**
22: Request *CLEM* to start assessing *out-LR* with the sink
23: Broadcast *FLPNICC*
24: **if** an *FLPNICC* Msg. is received • sender's *USN* is not me **then**
25: **if** my_status = *CM* • *TCM* **then**
26: Wait, for a defined time, to hear more *FLPNICC* from other neighbors
27: Evaluate *IC Comm*. offers
28: Request *CNMM* to mark the selected neighbor as your *LP USN*
Set *HTS* = selected neighbor's *HTS* + 1
29: **else if** my_status = *CH* **then**
30: Evaluate *LP IC Comm*. offers
31: Request *CNMM* to mark the selected neighbor as your *LP USN*
32: **else if** an *FLPNICC* Msg. is received • sender's *USN* is not me **then**
33: Request the *CNMM* to mark the sender as your *DSN*

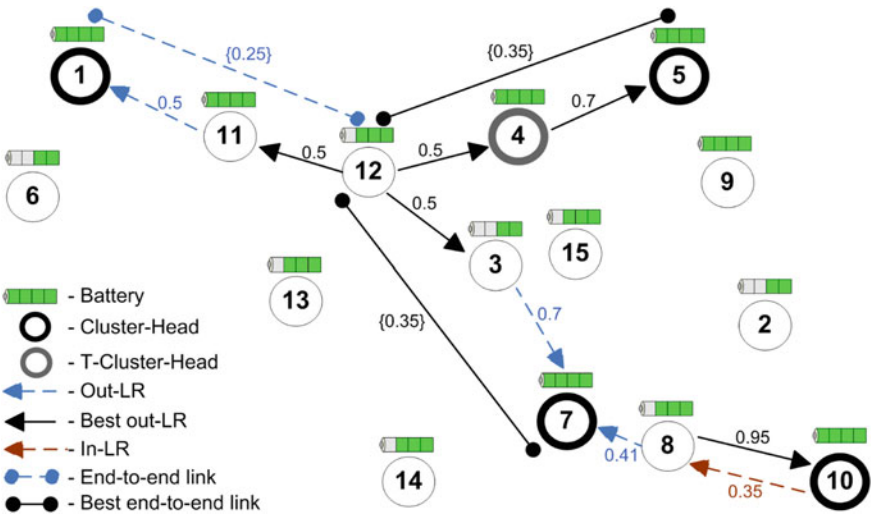


Fig. 11 Cluster formation process of *CCS*

Maintaining Network State Consistency

CCS makes use of the broadcast nature of the wireless communication to its advantage and exploits it to ensure network state consistency. For instance, after accepting a cluster joining offer, the node broadcasts either a Transitive Cluster Joining Notification (*TCJN*) or a Cluster Joining Notification (*CJN*), depending upon the type of cluster membership (direct or transitive), which includes the *NID* of the chosen *CH/TCH*. When a node receives *CJN/TCJN*, if it finds that it has been selected as a *CH/TCH*, it marks the sender as its *CM* or a sub-neighbor (*TCM*) in case of a transitive membership. If some other *CH/TCH* had the sender as its *CM* or *TCM* before, it un-marks the sender after finding out that it has joined some other cluster as a direct or transitive member. This approach exploits the broadcast nature of wireless communication and has a higher probability of keeping the network in a consistent state than the one where such notifications are unicast.

1.6 CEACH: Inter-Cluster Communication Overlay

For the sake of Inter-cluster Communication (*ICCOM*), the *CHs* form a multi-hop communication overlay in which communication takes place at high power. The process is started at the sink node, which broadcasts messages at high power that are meant for discovering *CHs* in its 1-hop range. The sink node broadcasts multiple messages to assess the link quality with the *CHs* in its 1-hop range. Upon hearing these messages, the *CHs* discover that they are at 1-hop from the sink node. They further broadcast this message at high power, with their hop count from the sink and the address of their upstream neighbor (which is the sink node itself for the *CHs* that are at 1-hop from it) along with number of times heard from the upstream neighbor, end-to-end in-link reliability with the sink, E_{RE} and direct membership count (number of direct *CMs*). Upstream neighbors overhear this forwarded message (kind of an implicit ACK (*IACK*)) and use information contained therein to assess its in- and out-Link reliability with the sender. The sink node and other upstream *CHs* send a Block ACK (*BACK*) to inform their downstream *CHs* of the number of times heard from them, information which is useful for them to compute their out-link reliability with their upstream *CHs*. Since it is very likely that the sink finds no *CH* in its 1-hop range, it broadcasts messages at low power also to discover ordinary nodes that lie in its low power 1-hop range. The same procedure is followed by the low power 1-hop neighbors of the sink except for that they send an *ACK* for each *ICCOM* discovery message that they hear from the sink. This *ACK* is used by the sink to assess its in-link reliability with these nodes. If a *CH* hears both high power as well as low power *ICCOM* discovery message sent by the sink, it uses low power to communicate with the latter instead of using high power. The sink also sends a *BACK* at low power to inform its 1-hop neighbors of the number of times it heard from each of them, information that is useful for these nodes to compute their out-link reliability with the sink.

These nodes forward the *ICCOM* discovery message further so that their *CHs* can discover routes back to the sink through them. In this way, these nodes serve as gateway nodes for *ICCOM* when either two *CHs* can't directly communicate at high power or a CH can't directly reach the sink node at high power. During the normal operation of the network, when *CHs* receive data from their *CMs* and aggregate it, if needed by the application, and relay it towards the sink using the *ICCOM*. If a CH finds no high power *CH* or the sink node in its 1-hop range, it uses one of these gateway nodes and forwards messages at low power to it. The *CHs* that lie at 1-hop low power range of the sink node also send the data to the sink at low power rather than sending it at high power. This adaptive power control contributes to conserve the battery of the nodes and helps prolong the life of the network.

1.7 Fuzzy-CEACH: Fuzzy Logic Based Cluster-Head Election

Fuzzy logic based description of the cluster head election process of Fuzzy-CEACH offers some interesting insights. The Fuzzy descriptors used for the cluster head election are the same as described in an earlier section, namely, node's residual energy, node's degree, node's in- and out-link reliabilities with its one-hop neighbors, node's hop-distance from a collection point/sink/or a central base-station.

These fuzzy descriptors serve as inputs to the fuzzy logic based cluster-head election process. This process, from the fuzzy logic perspective and following Mamdani's approach [26], consists of the following steps:

1.7.1 Fuzzification of the Crisp Inputs

The first step in the Fuzzy-CEACH is the conversion of the crisp inputs—residual energy, degree, in-/out-link reliability, hop-distance from the sink—into their respective fuzzy logic descriptors. These fuzzified descriptors are then allocated to their respective fuzzy sets. This process is popularly known as *fuzzification* in the fuzzy logic literature. This transformation involves assignment of crisp input variables to the linguistic descriptors used in fuzzy logic. For the fuzzy logic descriptor residual energy three fuzzy sets have been identified and they are *Low*, *Average*, and *High*. In Fig. 12 we depict the distribution of these fuzzy sets over the values of the fuzzy variable residual energy. Assuming a maximum residual energy that the wireless nodes could have is 5,000 mJ, the fuzzy set boundaries, which overlap and follow *trapezoidal distribution*, are shown in Fig. 12.

A node having a residual energy of 1,000 mJ, has a partial membership of the fuzzy set "*Average*" with a degree of membership 0.2. At the same time it has a

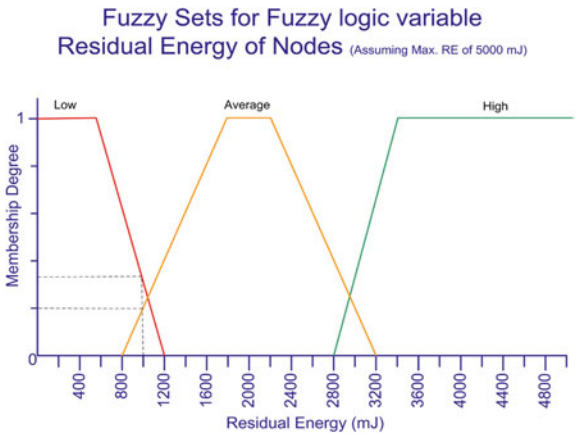


Fig. 12 Fuzzy sets for the fuzzy logic descriptor residual energy of the nodes (assuming a maximum residual energy of 5,000 mJ)

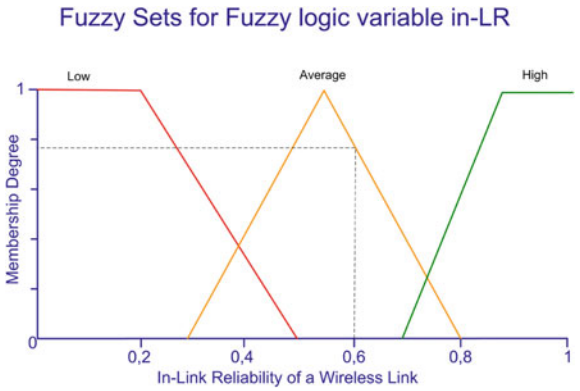


Fig. 13 Fuzzy sets for the fuzzy logic descriptor in-link reliability of a wireless link of a node

partial membership of the fuzzy set "*Low*" and its degree of membership in this set is 0.35.

Similarly, if the in-LR descriptor has a crisp value of 0.6, as shown in the Fig. 13, it will fall in the fuzzy set "*Average*" and its degree of membership will be 0.76.

Here, the degree of membership of the in-LR descriptor against a crisp value of 0.6 falls exclusively in the "*Average*" fuzzy set and is not a shared membership unlike in the previous example where residual energy had a crisp value 1,000 mJ which got a shared membership in two fuzzy sets, namely, "*Low"* and "*Average"*.

Figure 14 shows three fuzzy sets defined for the fuzzy logic descriptor Node Degree which is the third component considered in the Fuzzy-CEACH for choosing cluster heads. The fuzzy sets defined for this descriptor are also "*Low*", "*Average*", and "*High*". The node degree descriptor for a node with degree 6 belongs to the fuzzy set "*Average*" with degree of membership of 1. This again is a full membership, since the crisp value 6 is not shared among other two fuzzy

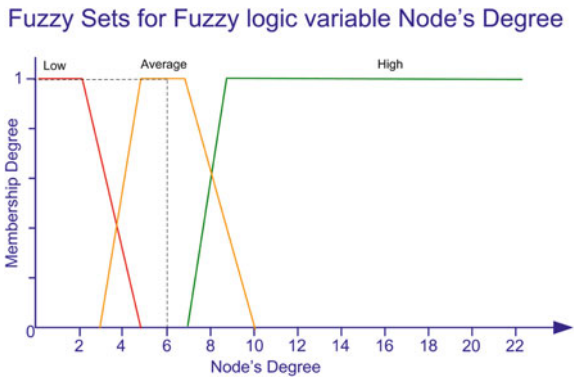


Fig. 14 Fuzzy sets for the fuzzy logic descriptor degree of a node

sets. Similarly, Figs. 15 and 16 show fuzzy sets for two fuzzy logic descriptors, namely, hop-distance of a node from the base station and the out-Link reliability of a wireless link of a node respectively.

1.7.2 Fuzzy Inference Engine

The core of a fuzzy system is the inference engine. It is this part where the fuzzified inputs are processed using the fuzzy Rule-Base and fuzzy reasoning process. For the cluster head election in Fuzzy-CEACH the Rule-Base used is given in Table 1. In order to keep things simple, we show only three fuzzified descriptors, namely, residual energy, node degree, and *MLR* in Table 1. The fourth fuzzified descriptor used in cluster head election in Fuzzy-CEACH is the hop-distance from the base-station. It is kept out of Table 1 to keep the table short and simple.

Generally, the linguistic variables used to represent fuzzy sets and their associated membership degrees for a given input are processed using simple functions like *MAX* (maximum) or *MIN* (minimum) depending upon the type of compound proposition in the antecedent. If in the compound proposition in the antecedent the individual statements are joined with a conjunction (and), the conjunction operator (AND, which means minimum) is applied to get a concluding value. If in the compound proposition in the antecedent the individual statements are joined with an OR, the disjunction operator (OR, which means maximum) is used to obtain a concluding value. For instance, the compound statement in the antecedent is a conjunction: "if the color of a fruit is orange and the shape is round, then it is an orange." On the other hand the compound statement, "if the service quality is bad or the food is not delicious or the food is very expensive, then the rating of the restaurant is *bad*", is a disjunction.

In Fuzzy-CEACH the application of the minimum or maximum functions is complemented by the importance that the application layer associates with each of these fuzzified variables. For instance, a normal fuzzy rule-base would simply apply either a *MAX* or a *MIN* function to get a resulting output fuzzy value.

Table 1 Rule-base for the fuzzy logic based cluster head election process

Rules-base for fuzzy logic based CH election				
Rule no.	RE	Node degree	MLR	$CHCV_{ELRout}(V_i)^*$
1	Low	Low	Low	Low
2	Low	Low	Average	Low
3	Low	Low	High	Average
4	Low	Average	Low	Low
5	Low	Average	Average	Low
6	Low	Average	High	Low
7	Low	High	Low	Low
8	Low	High	Average	Low
9	Low	High	High	Average
10	Average	Low	Low	Low
11	Average	Low	Average	Average
12	Average	Low	High	High
13	Average	Average	Low	Low
14	Average	Average	Average	Average
15	Average	Average	High	High
16	Average	High	Low	Low
17	Average	High	Average	Average
18	Average	High	High	Average
19	High	Low	Low	Average
20	High	Low	Average	High
21	High	Low	High	High
22	High	Average	Low	Low
23	High	Average	Average	High
24	High	Average	High	High
25	High	High	Low	Average
26	High	High	Average	Average
27	High	High	High	High

* The adaptive nature of Fuzzy-CEACH allows the allocation of importance factors to each fuzzy logic input descriptor, thus influencing the decision outcome

However, fuzzy-CEACH goes a step further in making the application layer's concerns addressed in terms of the preferences it associates with different fuzzy descriptors.

In Table 1, the value of the "*consequent*" fuzzy variable $CHCV_{MLR_{in}(V_i)}$ represents the chances of a node of being elected as a cluster head in its neighborhood. The value of this output variable ("*consequent*") varies between "*Low*" and "*High"*. For instance, rule number 12 says that "If a node has an "*Average"* residual energy and a "*Low*" degree and a "*High*" *MLR*, then it should have a "*High*" value of the output variable $CHCV_{MLR_{in}(V_i)}$". Since $CHCV_{MLR_{in}(V_i)}$ determines the likelihood of a node assuming the role of a cluster head, it is important to determine a crisp value associated with the suggested fuzzified output value of "*High*" as dictated by the rule number 12. This process of converting the output

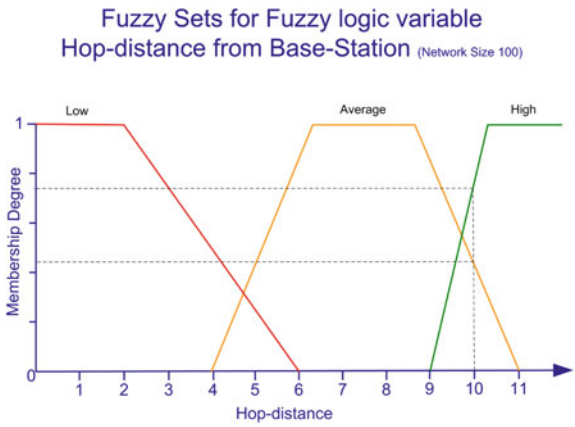


Fig. 15 Fuzzy sets for the fuzzy logic descriptor hop-distance from the base-station

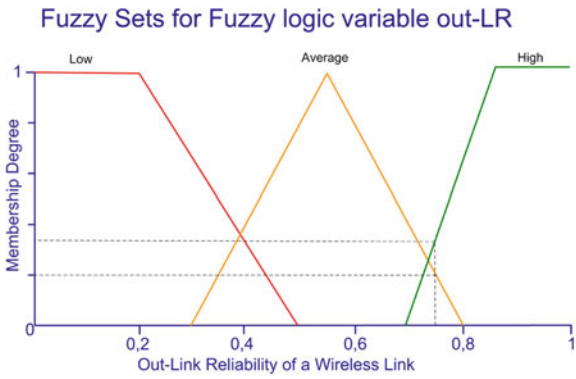


Fig. 16 Fuzzy sets for fuzzy logic descriptor out-link reliability of a wireless link of a node

value from fuzzified form to its corresponding crisp form is termed as defuzzification and is explained in the next section. The rule-bases for cluster formation and inter-cluster overlay formation process are given in Tables 2 and 3 respectively. We don't describe these two processes in detail here for the sake of brevity. These processes follow exactly the same pattern as the fuzzy-CEACH cluster head election process.

1.7.3 Defuzzification of the Output

The application of fuzzy reasoning through fuzzy inference engine produces a resulting fuzzy set which describes for each possible value how reasonable it will be to use it. In simpler terms one can say that for every possible value one knows the degree of membership to the resulting fuzzy set which indicates the reasonability of its use.

Table 2 Rule-base for the fuzzy logic based cluster formation process

Rule-base for fuzzy logic based cluster formation				
Rule no.	RE	Node degree	Out-LR	$CHCV_{ELRout}(V_i)^*$
1	Low	Low	Low	Low
2	Low	Low	Average	Low
3	Low	Low	High	Lowe
4	Low	Average	Low	Low
5	Low	Average	Average	Low
6	Low	Average	High	Low
7	Low	High	Low	Low
8	Low	High	Average	Low
9	Low	High	High	Low
10	Average	Low	Low	Low
11	Average	Low	Average	Average
12	Average	Low	High	High
13	Average	Average	Low	Low
14	Average	Average	Average	Average
15	Average	Average	High	High
16	Average	High	Low	Low
17	Average	High	Average	Average
18	Average	High	High	Average
19	High	Low	Low	Average
20	High	Low	Average	High
21	High	Low	High	High
22	High	Average	Low	Low
23	High	Average	Average	Average
24	High	Average	High	High
25	High	High	Low	Low
26	High	High	Average	Average
27	High	High	High	Average

* The adaptive nature of Fuzzy-CEACH allows the allocation of importance factors to each fuzzy logic input descriptor, thus influencing the decision outcome

However, in a fuzzy system one would like to get a crisp value as an output that the system applies or suggests to the user rather than a fuzzy set as an output. For instance, we don't want the system to tell us that the node has a high or low chance of being elected as a cluster head. Rather we want the system to tell us a numeric value for the $CHCV_{MLR_{in}(V_i)}$ fuzzy descriptor, so that we can determine if the node has a highest value in its one-hop neighborhood and if it is the node which will be elected as a cluster head. This process of transforming the output fuzzy set into a crisp value is called *defuzzification*.

Numerous approaches have been suggested in the fuzzy logic literature to achieve defuzzification. Some of the most popular ones are: center of gravity approach, center of singleton method, methods based upon finding maximum, margin properties of the centroid methods.

Table 3 Rule-base for the fuzzy logic based inter-cluster overlay formation process

Rule-base for fuzzy logic inter-cluster communication overlay formation

Rule no.	RE	Node degree	ELR-out	Hop-distance	$CHCV_{ELRout}(V_i)^*$
1	Low	Low	Low	Low	Low
2	Low	Low	Low	Average	Low
3	Low	Low	Low	High	Low
4	Low	Low	Average	Low	Average
5	Low	Low	Average	Average	Low
6	Low	Low	Average	High	Low
7	Low	Low	High	Low	Average
8	Low	Low	High	Average	Average
9	Low	Low	High	High	Low
10	Low	Average	Low	Low	Low
11	Low	Average	Low	Average	Low
12	Low	Average	Low	High	Low
13	Low	Average	Average	Low	Average
14	Low	Average	Average	Average	Average
15	Low	Average	Average	High	Low
16	Low	Average	High	Low	Average
17	Low	Average	High	Average	Average
18	Low	Average	High	High	Low
19	Low	High	Low	Low	Low
20	Low	High	Low	Average	Low
21	Low	High	Low	High	Low
22	Low	High	Average	Low	Average
23	Low	High	Average	Average	Low
24	Low	High	Average	High	Low
25	Low	High	High	Low	Average
26	Low	High	High	Average	Low
27	Low	High	High	High	Low
28	Average	Low	Low	Low	Low
29	Average	Low	Low	Average	Low
30	Average	Low	Low	High	Low
31	Average	Low	Average	Low	Average
32	Average	Low	Average	Average	Average
33	Average	Low	Average	High	Low
34	Average	Low	High	Low	Average
35	Average	Low	High	Average	Average
36	Average	Low	High	High	Average
37	Average	Average	Low	Low	Low
38	Average	Average	Low	Average	Low
39	Average	Average	Low	High	Average
40	Average	Average	Average	Low	Average
41	Average	Average	Average	Average	Average
42	Average	Average	Average	High	Low
43	Average	Average	High	Low	High

(continued)

Table 3 (continued)

Rule-base for fuzzy logic inter-cluster communication overlay formation

Rule no.	RE	Node degree	ELR-out	Hop-distance	$CHCV_{ELRout}(V_i)^*$
44	Average	Average	High	Average	High
45	Average	Average	High	High	Average
46	Average	High	Low	Low	Low
47	Average	High	Low	Average	Low
48	Average	High	Low	High	Low
49	Average	High	Average	Low	Average
50	Average	High	Average	Average	Average
51	Average	High	Average	High	Low
52	Average	High	High	Low	Average
53	Average	High	High	Average	Average
54	Average	High	High	High	Average
55	High	Low	Low	Low	Low
56	High	Low	Low	Average	Low
57	High	Low	Low	High	Low
58	High	Low	Average	Low	High
59	High	Low	Average	Average	Average
60	High	Low	Average	High	High
61	High	Low	High	Low	High
62	High	Low	High	Average	High
63	High	Low	High	High	Average
64	High	Average	Low	Low	Low
65	High	Average	Low	Average	Low
66	High	Average	Low	High	Low
67	High	Average	Average	Low	High
68	High	Average	Average	Average	High
69	High	Average	Average	High	Average
70	High	Average	High	Low	High
71	High	Average	High	Average	High
72	High	Average	High	High	High
73	High	High	Low	Low	Low
74	High	High	Low	Average	Low
75	High	High	Low	High	Low
76	High	High	Average	Low	Average
77	High	High	Average	Average	Average
78	High	High	Average	High	Average
79	High	High	High	Low	High
80	High	High	High	Average	High
81	High	High	High	High	Average

* The adaptive nature of Fuzzy-CEACH allows the allocation of importance factors to each fuzzy logic input descriptor, thus influencing the decision outcome

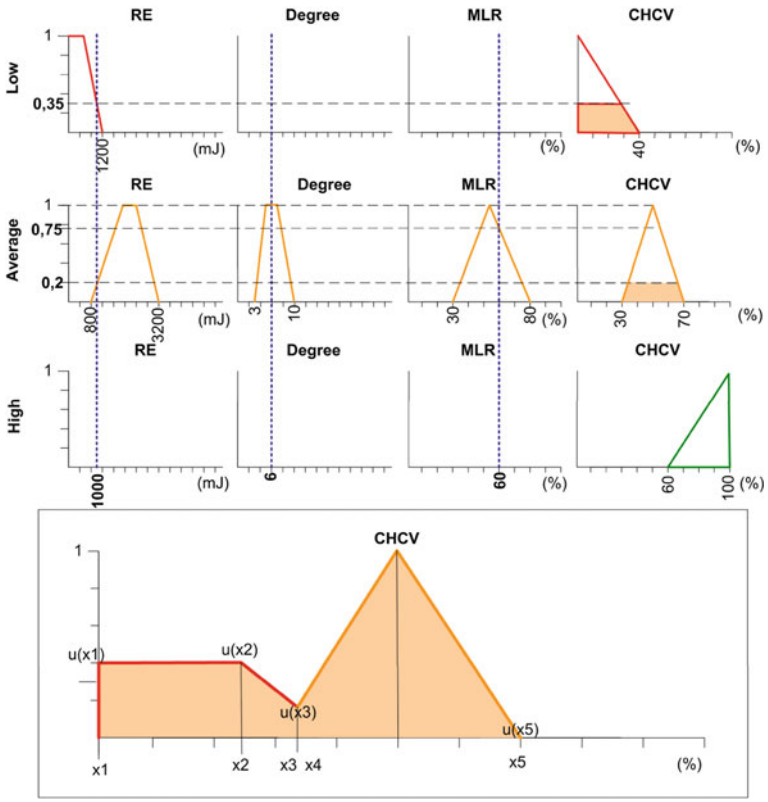


Fig. 17 Fuzzy-CEACH—fuzzy logic based CH election process showing different stages of fuzzy computations to obtain a crisp value for the output variable

In Fuzzy-CEACH we make use of the center of gravity approach to achieve defuzzification of the output fuzzy set. Figure 17 shows different stages of fuzzy computations to produce a fuzzy output set for $CHCV_{MLR_{in}(V_i)}$, whose crisp value according to the center of gravity approach comes out to be 30 %.

$$\text{Defuzzified } CHCV_{MLR_{in}(V_i)} = \frac{\sum_{i=1}^{5} x_i.u(x_i)}{\sum_{i=1}^{5} u(x_i)} = 30\ \%$$

1.8 Experimental Evaluation of CEACH, CTP and HEED

CEACH is implemented in Contiki [10] which is an open source operating system for programming low power WSN and other embedded systems used in the *Internet of Things*. Its programming model consists of multiple processes running

over an event driven kernel. It also supports multi-threading in the programs through an abstraction called protothreads. Other features include dynamically loadable modules, support for TCP/IP stack, Rime communication stack that offers protocol independent radio networking and a cross-layer network simulation tool called Cooja [11].

The performance of CEACH is compared against two well known protocols. One protocol is from the clustering domain called HEED and the other one is from link reliability based routing protocols domain called Collection Tree Protocol (CTP). The Contiki CTP implementation is taken for the comparisons presented in the subsequent sections.

Sample Application

In order to judge the relative performance of CEACH, HEED and the CTP against each other, a sample application is used that makes each node in the network send periodic data reports to the sink. The assumed period for these reports, called an *EPOCH*, is 14 s. In both CEACH and HEED protocols, the cluster heads upon receiving these data from their cluster members, apply desired data reduction functions, *MAX* in the present case, and forward the aggregated data in the upstream direction towards the sink through the clustering backbone. Since CTP is not a clustering protocol, it doesn't incorporate any data aggregation scheme. It, however, applies link reliability assessments in choosing more reliable routing paths.

Simulation Model

The results included in the subsequent sections are obtained using Cooja network simulator. The simulator is used to simulate WSN random deployments without any particular restrictions on the node density or distribution. The simulated nodes are tmote sky [16]. The link reliabilities are controlled programmatically. Having different reliabilities on different links help one assess the performance of a protocol. If a given clustering protocol or a routing protocol does consider link reliabilities, it should base its decisions on the assessed reliability values on different links. This methodology is used to assess the performance of CEACH, HEED and the CTP. Different network sizes are used to see the behavior of the protocols when the network scales. The energy consumed by the nodes is assessed using the power profiling mechanism [10] provided by Contiki. The same initial energy of 1,000 mJ is assigned to each node in the experiments, unless otherwise stated. The Contiki energy profiling framework measures times for which different components of the nodes remain active. This information, along with current consumption from tmote sky data-sheet, is used to compute the energy consumption of a node.

Data Delivery Ratios

The primary objective for incorporating link reliability in route selection is to ensure the selection of those communication paths that could transport data reliably to the sink. For this purpose the CEACH approach incorporates different incarnations of link reliability through different phases of its operation. In order to

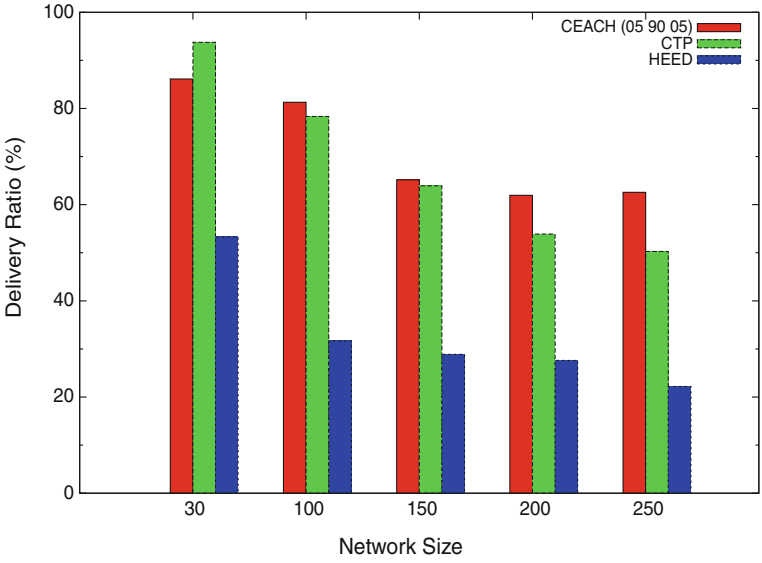


Fig. 18 Delivery ratio at the sink node for CEACH, CTP and HEED

measure the performance of CEACH, HEED and CTP on reliability count, Data Delivery Ratio (*DDR*) at the sink node is measured and the results are presented here. Data delivery ratio, in principle, is the ratio of the number of messages that are successfully received at a destination node, which is the sink in this case, to the total number of messages that are generated by the network. The data delivery ratios achieved by CEACH, HEED and the CTP are measured for different network sizes and the results are presented here. For the CEACH protocol the value of the *Impact Factors* (*IFs*) is also varied in the *CHCV* metric to assess its impact on the data delivery ratios. The results are plotted in Fig. 18.

The different combinations of the values of the used *Impact Factors* (in percent) are also shown in the legends part of the graph. They appear in the order IF_{REI}, $IF_{MLR_{in}}$ and IF_{NDI}. It is evident from this plot that CEACH does achieve high data delivery ratios when appropriately high *Impact Factor* is assigned to the link reliability parameter in the *CHCV* metric. Additionally, the data delivery ratio is not adversely affected even when the network scales. The slight downward trend that one observes in the graph is due to different network dynamics in each deployment. It should be noted here that no explicit retransmissions are used and each node just sends a message only once to its cluster head or *TCH*. The data delivery ratios achieved by HEED and CTP are lower than that of CECAH in all network sizes. This is because of the fact that HEED protocol remains completely agnostic to the link quality in the clustering process and any good or bad communication paths chosen by it are purely incidental.

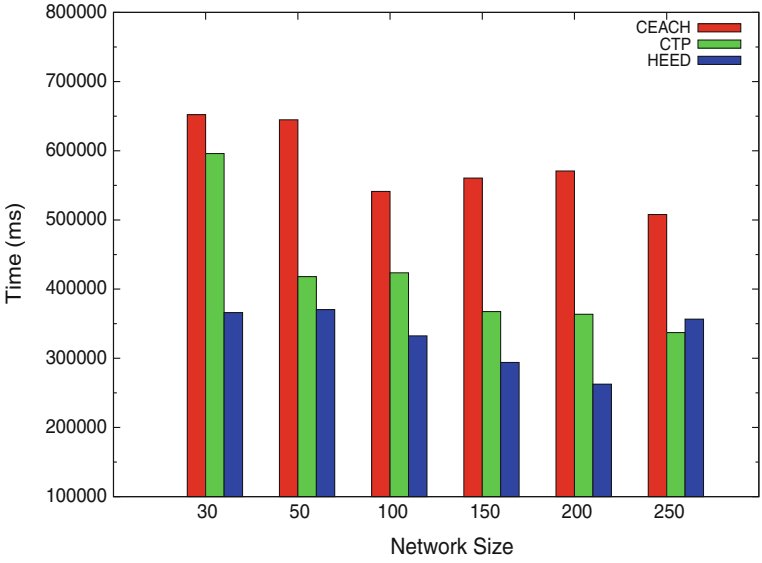


Fig. 19 Network lifetime depicted for CEACH, CTP and HEED in terms of first node died (*FND*)

Network Lifetime

Several definitions of network lifetime have been suggested in the WSN and other ad-hoc wireless networks literature. There has been no consensus on one definition that could serve all application scenarios. This section describes the most widely used definitions by the WSN and wireless ad-hoc networks community along with their advantages and disadvantages.

The performance of the protocols is measured not only through data delivery ratios at the sink but also by measuring the time, in milliseconds, at which major of the nodes die making the network disconnected. The performances of CEACH, HEED, and the CTP are plotted and shown in Figs. 19, 20 and 21.

First Node Dies (*FND*)

Some approaches measure the network lifetime as the duration of time between the start of network operations to the instance of time when the first node dies. This approach to measuring network lifetime, however, overlooks the basic purpose for which the network is there in the first place, i.e. to sense and send data towards the sink. Since these tasks could be carried on by the rest of the nodes after the death of the first node, this approach fails to account for this simple fact. However, it can still be used to characterize the protocols on their balanced energy consumption in the network. Since the nodes which consume energy at a higher rate in the network die first, the early death of a node may indicate a potential unbalanced use of the energy in the network.

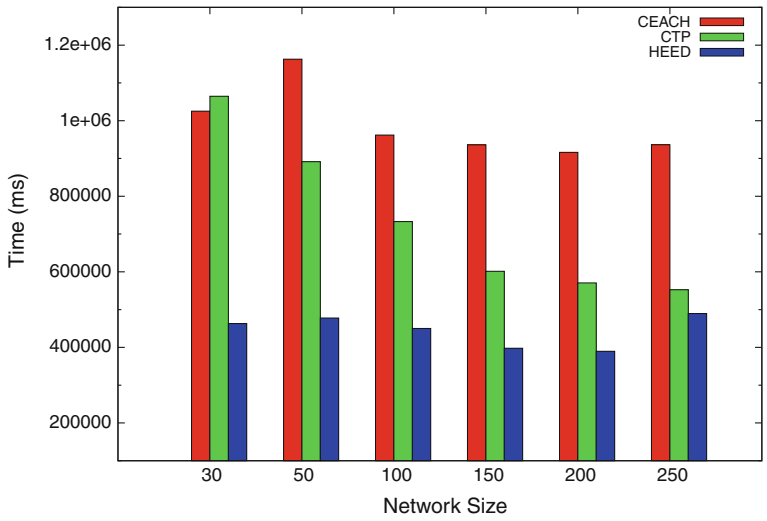


Fig. 20 Network lifetime in terms of half of the nodes died (*HND*) for CEACH, CTP, and HEED

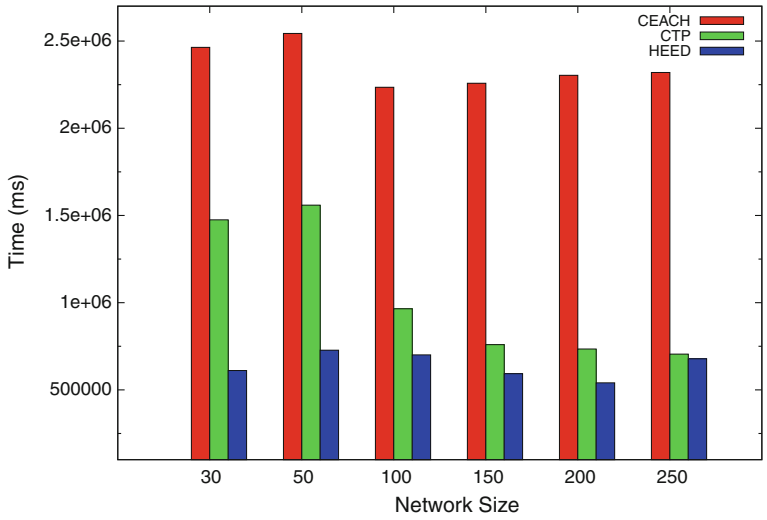


Fig. 21 Network lifetime in terms of majority of the nodes died (*MND*) for CEACH, CTP, and HEED

Plot in Fig. 19 shows the network lifetime, in terms of *FND*, for CEACH, HEED, and CTP. CEACH does a better job on all network sizes than HEED and CTP, because of its balanced energy usage in the network and its usage of link reliability assessments in choosing *CHs* as well as upstream inter-cluster routing paths. Data reduction functions used in CEACH also help lower the network

traffic, thus prolonging the network lifetime. CTP also does a good job in smaller network sizes. However, as the network scales its performance degrades for not using any data aggregation scheme.

Half of the Nodes Die *(HND)*

Another commonly used definition of Network lifetime states that it is the duration of time between the start of the network operations to the instance of time when half of the nodes in the network die. One apparent problem with this approach is that it is not that convenient to determine if half of the nodes in the network have actually died. It is normally the sink node that is responsible for determining how many nodes have died and it does that on the basis of the received data/control messages. This implies that there needs to be mechanisms in place for the sink node to keep an account of all the nodes in the network and all nodes have to continuously provide their residual energy status to the sink node. Additionally, if the remaining half of the nodes in the network can still carry on their work and provide the necessary coverage and successfully meet the fidelity constraints set by the application, then the network should not be termed as dead.

On a positive side, *HND* metric does provide a basis of comparison between networking protocols for wireless ad-hoc networks. The longer it takes, under a given protocol, for the network to reach this milestone, the better is the protocol and vice versa. Additionally in a simulated environment, it is relatively easy to measure the network life time using this metric. Figure 20 shows the plot of the *HND* metric for CEACH, CTP, and HEED. On smaller network sizes both CEACH and CTP does well, whereas the HEED protocol performs poorly. One possible reason for HEED to perform badly is the fact that it doesn't incorporate link reliability in choosing cluster heads and inter-cluster routing paths. As a result the nodes spend higher energies in repetitive transmissions of the same message, thus depleting them quickly. The clustering process of HEED also takes longer and thus consumes more energy.

On smaller network sizes, like networks consisting of 30 odd nodes, CTP does well in terms of *HND* metric. However, as the network scales, CTP's performance tends to suffer. One apparent reason for CTP to do well on small network sizes is the fact that most, if not all, nodes can directly communicate with the sink node and there are no parent nodes to be chosen and the protocol overhead in terms of beacon messages doesn't show up negatively on its performance.

Majority of the Nodes Die (*MND*)

This metric regards the network lifetime as the duration of time from the beginning of the network operations to the instance of time when majority of the nodes die, thus making the network disconnected and rendering it unable to carry out its data communication tasks. This metric doesn't have the shortcomings that mar the other metrics, namely, *FND* and *HND*. Although finding an agreement over the definition of what constitutes a majority of nodes in a wireless ad-hoc network like a WSN, a WSAN, or a MANET is difficult, still this approach is most meaningful for defining the network lifetime. One plausible definition of majority nodes is that

when the sink node stops receiving data/control messages for a specific period of time. This could mark the death of majority of the nodes in the network and that the network cannot sustain its data communication tasks. This definition of the network lifetime is very plausible and is used to plot the results presented here in this section. The results presented below in Fig. 21 assume the network as dead when the sink node doesn't receive any data reports for successive 10 *EPOCHS*.

Again the CEACH approach shows better results, for *MND* metric, HEED and CTP. The real effect of balanced energy usage in the network is reflected in this plot. CTP does a better job than HEED but it doesn't come close to the CEACH approach on *MND* metric. The reason for this behavior lies in adaptive beaconing that is used in CTP for link reliability assessments. As more nodes start dying due to running low on battery, the remaining ones start reducing their beaconing interval, since the ETX values on the links start getting higher, because the dead nodes can not send an acknowledgement or a beacon. This makes the nodes that are yet alive to reduce the beaconing interval even further ultimately causing a quick battery discharge.

2 CEACH for Distributed Event-Based Systems

CEACH middleware service could be applied in WSAN for quite a few scenarios ranging from spatio-temporal in-network aggregation to energy efficient hierarchical routing. One promising application of CEACH in distributed event-based systems can be to manage and control the distribution of event notifications. Since publish-subscribe mechanisms have been found very practical for distributing event notifications from producers to the interested consumers. The CEACH approach suggests cluster head nodes in the network which are more capable in terms of the attributes that are important for a given application. These nodes can serve as event notification brokers (notification routers in Fig. 22) and can manage the distribution of event notifications to the event consumers as and when they receive a published event by an event producing node in a cluster. If both event producer and consumer are part of the same cluster, then for an event broker the delivery of the notification is local to the cluster.

However, if the event consumers are spread over multiple clusters, then a federated system of brokers would be needed to distribute the event notifications. Such a federated broker network could be generated using the inter-cluster communication offered by the CEACH approach. Considering scarce resources in a WSAN, having global knowledge at each cluster head of all subscriptions is not feasible. Therefore, each cluster head could just manage subscriptions that belong to its own cluster.

If a published event doesn't have a relevant subscription in the same cluster, the cluster head can broadcast this event notification at high power to its peers in the federated overlay of cluster heads. Since only a subset of the nodes (cluster heads) is involved in forwarding such event notifications, the overall communication costs can be kept low.

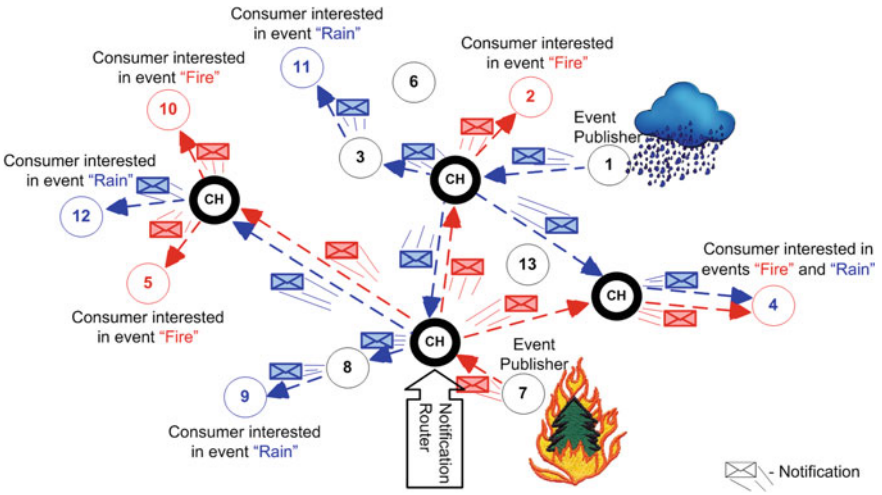


Fig. 22 Application of CEACH middleware service in a publish/subscribe based distributed event based system

Figure 22 shows two event publishing nodes (*node*-1 and *node*-7). These nodes detect different events, in the present case these are the event '*Rain*' and the event '*Fire*'. The event '*Rain*' is detected by node-1 and it publishes a notification for this event and sends it to its *CH*. The *CH* keeps the corresponding event subscriptions from the nodes within its cluster. There is only *node*-11 belonging to the same cluster which is interested in the event '*Rain*' and is, therefore, promptly informed of the event by the *CH* via node-3. The *node*-11 might be attached with an actuator to cause appropriate actuation for the event '*Rain*', probably turning the water sprinklers off in a garden as the plants don't need watering when it rains. This event notification is also broadcast by the *CH* in the cluster head overlay at high power, as there might be other nodes belonging to other clusters which are interested in the event '*Rain*'. The *CHs* receiving the '*Rain*' event notification check their subscription lists of their members and if there is a member node interested in this event, it is promptly informed of the event through the event notification. In the present case two nodes (9 and 12) belonging to different clusters receive the event '*Rain*' notification. The other event '*Fire*' is handled in a similar manner in the network.

3 Conclusions

The chapter described important issues faced by cyber-physical systems, in general, and WSN and WSAN in particular. Some important application areas of cyber-physical systems have also been described. For Cyber-Physical systems to

achieve wide-spread acceptance and success, the issues like reliable communication of events between different wirelessly communicating nodes of the system, application- and channel-awareness and energy-efficiency have to be addressed. The CEACH approach, described in this chapter, makes few strides towards achieving these goals. The initial results are encouraging and they show the promise that the CEACH approach offers for Cyber Physical Systems. However, there is a need to investigate the suitability of the CEACH approach in more challenging application scenarios. One such area is distributed event based systems where event notification distribution can be achieved using the CEACH approach. However, this area also needs to be further investigated and the suitability of the CEACH approach in this area needs to be established further by experimental results. The chapter also described Fuzzy-CEACH which is an extension of the CEACH approach and uses fuzzy logic principles.

References

1. I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci, Wireless sensor networks: a survey. Comput. Netw. **38**, 393–422 (2002)
2. AS-International Association e.V., Actuator-sensor interface (2013), http://www.as-interface.net/. Accessed 10 June 2013
3. D. Baker, A. Ephremides, The architectural organization of a mobile radio network via a distributed algorithm. IEEE Trans. Commun. **29**, 1694–1701 (1981)
4. M. Buettner, G.V. Yee, E. Anderson, R. Han, X-MAC: a short preamble MAC protocol for duty-cycled wireless sensor networks. in *Proceedings of the 4th International Conference on Embedded Networked Sensor Systems (SenSys '06)*. ACM, New York, USA, 2006, pp. 307–320
5. CAN in Automation e.V., Kontumazgarten 3. DE-90429 Nuremberg (2013), http://www.can-cia.org/. Accessed 09 June 2013
6. M. Chatterjee, S.K. Das, D. Turgut, WCA: a weighted clustering algorithm for mobile ad hoc networks. J. Cluster Comput. (Special Issue on Mobile Ad hoc Networks), **5**, 193–204 (2002)
7. A. Cerpa, J.L. Wong, M. Potkonjak, D. Estrin, Temporal properties of low power wireless links: modeling and implications on multi-hop routing. in *MobiHoc'05*, Urbana-Champaign, IL, May 2005
8. D.S.J. De Couto, D. Aguayo, J. Bicket, R. Morris, A high-throughput path metric for multi-hop wireless routing. in *Proceedings of the 9th annual international conference on Mobile computing and networking (MobiCom '03)*. ACM, New York, USA, 2003, pp. 134–146
9. A. Dunkels, B. Grönvall, T. Voigt, Contiki—a light-weight and flexible operating system for tiny networked sensors, in *First IEEE Workshop on Embedded Netwoked Sensors (Emnets-I)*, 2004
10. A. Dunkels, F. Österlind, N. Tsiftes, Z. He, Software based online energy estimation for sensor nodes, in *4th IEEE Workshop on Embedded Netwoked Sensors (Emnets-IV)*, 2007
11. J. Eriksson, F. Österlind, N. Finne, N. Tsiftes, A. Dunkels, T. Voigt, R. Sauter, P.J. Marrón, COOJA/MSPSim: interoperability testing for wireless sensor networks. in *Proceedings of the 2nd International Conference on Simulation Tools and Techniques (Simutools '09). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering)*, Brussels, Belgium, 2009

12. O. Gnawali, R. Fonseca, K. Jamieson, D. Moss, P. Levis, Collection tree protocol. in *Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems (SenSys '09)*. ACM, New York, USA, 2009, pp. 1–14
13. J. Goldman, K. Shilton, J. Burke, D. Estrin, M. Hansen, N. Ramanathan, S. Reddy, V. Samanta, M. Srivastava, R. West, Participatory sensing: a citizen-powered approach to illuminating the patterns that shape our world Foresight & Governance Project, White paper, 2009
14. W.R. Heinzelman, A. Chandrakasan, H. Balakrishnan, Energy-efficient communication protocol for wireless microsensor networks, in *HICSS'00: Proceedings of the 33rd Hawaii International Conference on System Sciences-Volume 8*, IEEE Computer Society, Washington, DC, USA, 2000, p. 8020
15. D. Jacobi, P.E. Guerrero, K. Nawaz, C. Seeger, A. Herzog, K. Van Laerhoven, I. Petrov I, *Towards Declarative Query Scoping in Sensor Networks*. ed. by K. Sachs, I. Petrov, P. Guerrero. From Active Data Management to Event-Based Systems and More, Lecture Notes in Computer Science 6462, ISBN 978-3-642-17225-0, (Springer, 2010)
16. M. Johnson, M. Healy, P. van de Ven, M.J. Hayes, J. Nelson, T. Newe, E. Lewis, A comparative review of wireless sensor network mote technologies. in *IEEE Sensors 2009*, 2009, pp. 1439–1442
17. H. Kang, In-network processing of joins in wireless sensor networks. Sensors **13**, 3358–3393 (2013)
18. K. Katsalis, A. Xenakis, P. Kikiras, G. Stamoulis, Topology optimization in wireless sensor networks for precision agriculture applications. in *Proceedings of the 2007 International Conference on Sensor Technologies and Applications (SENSORCOMM '07)*. IEEE Computer Society, Washington, DC, USA, 2007, pp. 526–530
19. S. Kim, S. Pakzad, D. Culler, J. Demmel, G. Fenves, S. Glaser, M. Turon, Health monitoring of civil infrastructures using wireless sensor networks, in *Proceedings of the 6th International Conference on Information Processing in Sensor Networks*, *IPSN'07*, ACM, New York, USA, 2007, pp. 254–263
20. M. Kovatsch, S. Duquennoy, A. Dunkels, A low-power CoAP for Contiki. in *Proceedings of the 2011 IEEE Eighth International Conference on Mobile Ad-Hoc and Sensor Systems (MASS '11)*. IEEE Computer Society, Washington, DC, USA, 2011, pp. 855–860
21. B. Krishnamachari, D. Estrin, S. Wicker, The impact of data aggregation in wireless sensor networks. in *Proceedings of the 22nd International Conference on Distributed Computing Systems (ICDCSW '02)*. IEEE Computer Society, Washington, DC, USA, 2002, pp. 575–578
22. J.K.S Lau, C.K. Tham, T. Luo, Participatory cyber physical system in public transport application. in *Fourth IEEE International Conference on Utility and Cloud Computing*, ucc, pp. 355–360
23. T. Liu, A. Kamthe, L. Jiang, A. Cerpa, Performance evaluation of link quality estimation metrics for static multihop wireless sensor networks, sensor, mesh and ad hoc communications and networks, 2009. SECON '09. in *6th Annual IEEE Communications Society Conference on*, pp. 1–9, 22–26
24. S.R. Madden, M.J. Franklin, J.M. Hellerstein, W. Hong, TinyDB: an acquisitional query processing system for sensor networks. ACM Trans. Database Syst. **30**, 1 (2005)
25. A. Mainwaring, D. Culler, J. Polastre, R. Szewczyk, J. Anderson, Wireless sensor networks for habitat monitoring. in *Proceedings of the 1st ACM International Workshop on Wireless Sensor Networks and Applications (WSNA '02)*. ACM, New York, USA, 2002, pp. 88–97
26. E.H. Mamdani, Application of fuzzy algorithms for the control of a simple dynamic plant. in *Proceedings of IEEE*, 1974, pp. 121–159
27. B. McMillin, C. Gill, M.L. Crow, F. Liu, D. Niehaus, A. Potthast, D. Tauritz, Cyber-physical systems distributed control: the advanced electric power grid. in *Beyond SCADA*, Nov 2006
28. A. Milenkovi′c, C. Otto, E. Jovanov, Wireless sensor networks for personal health monitoring: issues and an implementation. Comput. Commun. (Special issue: Wireless Sensor Networks: Performance, Reliability, Security, and Beyond) **29**, 2521–2533 (2006)

29. K. Nawaz, A. Buchmann, ACDMCP: an adaptive and completely distributed multi-hop clustering protocol for wireless sensor networks. Int. J. Wireless Mob. Networks. No. ISSN: 0975–3834, **2**, 0975–4679 (2010) AIRC
30. K. Nawaz, A.P. Buchmann, Building a case for FIPA compliant multi- agent based approaches for wireless sensor networks. in *Proceedings of 1st ICST International Conference on Ambient Media and Systems*. Software Organization and Monitoring of Ambient Systems Workshop, ACM
31. K. Nawaz, P.E. Guerrero, A.P. Buchmann, Towards a FIPA compliant multiagent based middleware architecture for sensor networks. in *Proceedings of 3rd IEEE International Symposium on Wireless Pervasive Computing (ISWPC 2008)*, *IEEE*, 2008
32. Y. Sun, B. McMillin, X. Liu, D. Cape, Verifying noninterference in a cyber-physical system. The advanced electric power grid. in *Proceedings of the Seventh International Conference on Quality Software (QSIC '07)*. IEEE Computer Society, Washington, DC, USA, 2007, pp. 363–369
33. B. Sundararaman, U. Buy, A.D. Kshemkalyani, Clock synchronization for wireless sensor networks: a survey, ad-hoc networks. Elsevier **3**, 281–323 (2005)
34. A.M. Tabar, A. Keshavarz, H. Aghajan, Smart home care network using sensor fusion and distributed vision-based reasoning. in *VSSN'06: Proceedings of the 4th ACM International Workshop on Video Surveillance and Sensor Networks*, ACM, New York, USA, pp. 145–154
35. The CeNSE Project, Central nervous system for the earth by Hewlett Packard Development Company, L.P. (2013), http://www8.hp.com/us/en/hp-information/environment/cense.html. Accessed 27 March 2013
36. R. Virrankoski, A. Savvides, TASC: topology adaptive spatial clustering for sensor networks, in *in IEEE International Conference on Mobile Adhoc and Sensor Systems MASS 2005*, 2005, pp. 605–614
37. J. Wang, H. Abid, S. Lee, L. Shu, F. Xia, A secured health care application architecture for cyber-physical systems. Control Eng. Appl. Inform. **13**(3), 101–108 (2011)
38. D. Work, A. Bayen, Q. Jacobson, Automotive cyber physical systems in the context of human mobility. in *Proceedings of the National Workshop on High-Confidence Automotive Cyber-Physical Systems* Troy, MI, USA, 2008
39. F. Xia, A.V. Vinel, R. Gao, L. Wang, T. Qiu, Evaluating IEEE 802.15.4 for cyber-physical systems. EURASIP J. Wireless Comm. Networking, 2011
40. Y. Yao, J. Gehrke, The cougar approach to in-network query processing in sensor networks. SIGMOD Rec. **31**(3), 9–18 (2002)
41. O. Younis, S. Fahmy, HEED: a hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks. IEEE Trans. Mob. Comput. **3**, 366–379 (2004)
42. L. Zadeh, Fuzzy sets. Inf. Control **8**, 338–353 (1965)
43. N.M. Zarmehri, A. Aguiar, Supporting sensing application in vehicular networks. in *ACM MobiCom Workshop on Challenged Networks*, 2012

Multi-objective Optimization for Error Compensation in Intelligent Micro-factory CPS

Azfar Khalid and Zeashan H. Khan

Abstract In the last decade, the demand of micro products and miniaturization has seen a wide spread growth. Currently, micro products and micro features are produced through conventional macro scale ultra-precision machines and MEMS manufacturing techniques. These technologies have limitations as conventional machining centers consume large energy and space. For mass production of micro components using non-silicon materials and real 3D shapes or free-form surfaces, mechanical micro manufacturing technology based machine tools are developed as an alternative method. The principle of "Small equipment for small parts" is gaining trend towards the investigation on micro-machine tools. One example of miniaturization of manufacturing equipment and systems is the Japanese micro-factory concept. Few micro-machines and associated handling micro grippers and transfer arms are developed to create micro-factory. The manufacturing processes are performed in a desktop factory environment. To explore the micro-factory idea, large number of micro machines can be installed in a small work-floor. The control of this micro factory concept for operation, maintenance and monitoring becomes a Cyber-physical system capable of producing micro-precision products in a fully-automated manner at low cost. Manufacturing processing data and condition monitoring of micro machine tools in a micro factory are the variables of interest to run a smooth process flow. Every machine out of hundreds of micro machines will have sensing equipment and the sensors data is being compiled at one place, ideally using wireless communication systems. One or two operators can run and monitor the whole micro-factory and access the machine if the fault alarms receive from any station. A variety of sensors will be employed for machine control, process control, metrology and calibration, condition monitoring

A. Khalid (✉)
Department of Mechanical Engineering, Mohammad Ali Jinnah University (MAJU), Islamabad, Pakistan
e-mail: azfar.khalid@jinnah.edu.pk
URL: http://www.jinnah.edu.pk

Z. H. Khan
National University of Science and Technology, Islamabad, Pakistan

Z. H. Khan et al. (eds.), *Computational Intelligence for Decision Support in Cyber-Physical Systems*, Studies in Computational Intelligence 540,
DOI: 10.1007/978-981-4585-36-1_2,

of machine tools, assembly and integration technology at the micro-scale resulting in smooth operation of micro-factory. Single machine can be designed with a computer numerical control, but, flexible reconfigurable controllers are envisioned to control variety of processes that will lead to the development of open architecture controllers to operate micro-factory. Therefore, the control effort and algorithms have to utilize process models to improve the overall process and, ultimately, the product. Thus, we aim to introduce machine to machine (M2M) communication in the micro factory test bed. M2M communication enables micro actuator/sensor & controller devices to communicate with each other directly i.e., without human intervention, automating management, monitoring, and data collection between devices, as well as communicating with neighboring machines. All micro sensors communicate with a local short distance wireless network e.g. via Bluetooth piconet as well as with a centralized controller via WLAN 802.11 to exchange control/command from it. In this chapter, inherent issues are first highlighted where bulk micro-part manufacturing is carried out using large size machines. State-of-the-art micro machine tool systems designed and developed so far are discussed. With the help of precision engineering fundamentals and miniaturization scaling issues, a design strategy is formulated for a high precision 3-axis CNC micro machine tool as a model for micro-factory working. Based on this, a mathematical model is built that includes machine's design variables and its inherent errors. The volumetric error between tool/work-piece is evaluated from the machine's mathematical model and further used as an objective function to be minimized. Robust design optimization at micro machine development stage reveals the sensitivity analysis of each design variable. The optimization analysis employs different design of Experiment (DOE) techniques to make initial population that is governed by multi-objective genetic algorithm. Hence, the robust design is achieved for 3-axis micro machine tool using the essential knowledge base. The technique is used to remove the machine's repeatable scale errors via calibration and is known as error mapping. These errors are entered into the machine controller, which has the capability of compensating for the error. The machine does not need any extra hardware. Error mapping is a cost-effective tool in achieving volumetric accuracy in a micro manufacturing system.

Keywords Micro factory · Micro machines · Robotic cyber physical system · Machine to machine communication · Volumetric error · Error compensation

1 Introduction

In the last decade, the demand of micro products and miniaturization has seen a wide spread growth. The application areas for miniaturized products are micro-sensors, accelerometers, micro-mirrors, fiber optics connectors and micro and nano electronics components. Only the IT peripheral and biomedical industries have

consumed micro products of worth US$80 billion in 2010. Currently, micro products and micro features are produced through conventional macro scale ultra-precision machines and MEMS manufacturing techniques. These technologies have limitations as conventional machining centers consume large energy and space. Some machine tool manufacturers equip machining centers with costly micro and nano scale machining accessories which caters meso and micro products manufacturing. On the other hand, MEMS is photolithography based fabrication technique. Although, it is an efficient technology for the mass production of micro parts but is limited in terms of variety of raw materials and 2D structures.

For mass production of micro components using non-silicon materials and real 3D shapes or free-form surfaces, mechanical micro manufacturing technology based machine tools are developed as an alternative method. The principle of "Small equipment for small parts" is gaining trend towards the investigation on micro-machine tools, such as, micro-lathe, micro-milling, micro-press and micro components holding, assembling and transfer devices. One example of miniaturization of manufacturing equipment and systems is the Japanese micro-factory concept. Few micro-machines and associated handling micro grippers and transfer arms are developed to create micro-factory. The manufacturing processes are performed in a desktop factory environment.

The meso scale lies between the molecular or atomistic scale (where it is convenient to describe molecules in terms of a collection of bonded atoms) and the continuum or macroscale (where it is convenient to describe systems as continuous). The idea of desktop machines was initiated for the concept of 'micro factories' in the previous decade. Presently, the meso scale parts are manufactured with various processes like electrolytic in-line dressing (ELID), Electro-chemical machining (ECM), die sinking electro discharge machining (EDM), wire cut electro discharge machining (WEDM), milling, turning etc. However, lithography based techniques are the most common for micro manufacturing. As the size of the products become increasingly smaller and market demand for meso scale parts are on the increase, the previous non-lithographic processes are required to be employed at the micro and meso scale. Many researchers in Japan have already developed micro machines that can be mounted on a table top.

This chapter first classifies the non-lithography based micro manufacturing on the basis of machine tool size. Several instances are provided for the use of both standard and desktop size machines in micro manufacturing. Several benefits are identified for the use of miniaturized machines in micro manufacturing processes. Challenges and foreseeable problems for the design and control of desktop size machines and micro-factories are discussed. A design strategy is proposed based on the robust design and optimization technique for a desktop machine tool. The technique can be extended to the design of a micro-factory.

1.1 Non-lithography Based Classical Micro-manufacturing

Non-lithography based meso scale parts are manufactured with different processes and from different size of machines ranging from small to large scale. Classification can be made in non-lithography based meso scale parts manufacturing, based on the size of machines employed for the purpose. Currently, two major groups exist for non-lithography based micro manufacturing; small scale machines often called desktop machines and the standard size machines. Size of machine is identified based on the total volume of the machine tool. Standard size machine tool developers have equipped standard machines for micro machining. Many researchers [1, 2] have also used standard machines for the fabrication of meso scale parts. Some commercial state of the art machine tools equipped with micro manufacturing modules are discussed below.

AGIE [3] **(Switzerland)**: This manufacturer is a leading supplier of ultra precision Wire Cut EDM and Die Sinking EDM. In addition to the different model series for WEDM like AGIECUT VERTEX, AGIECUT CLASSIC and for Die sinking EDM like AGIETRON SPIRIT and AGIETRON HYPERSPARK, it also offers EDM for micro and nano scale application. "Agietron Micro-Nano" is a die sinking machine that can perform micro drilling to machining of micro structures with the addition of nano module. Machine's micro module has a travel range of 220/160/100 mm^3 (*X/Y/Z*), positioning accuracy of ± 1 μm, resolution, 0.1 μm and the surface roughness of work piece, Ra 0.1 μm. The nano module has a travel range of 6/6/4 mm^3 (*X/Y/Z*) and can fit on the same machine by replacing the rotary axis of the AGIETRON micro. The nano module uses voice coil linear motors and can achieve a positioning accuracy of ± 0.1 μm, resolution 0.02 μm and the surface roughness of work piece, Ra 0.05 μm. Nano module has used parallel kinematics for axes movements (See Fig. 1) .

PRIMACON [4] **(Germany)**: It has developed a 5-axis vertical machining centre (Model # PFM 4024-5D) for the manufacturing of small components. The machine has a travel range of 400/240/350 mm^3 (*X/Y/Z*) with a positioning accuracy, under 1 μm and a rotational repeatability of 1 s arc. The machine uses Heidenhain iTNC 530 controller for the CNC.

FANUC [5] **(Japan)**: The model of this company for micro manufacturing is ROBONANO α-0iB (See Fig. 2), 5 axis CNC precision machining centre. It is a multi-purpose machine used for milling, turning, grinding and shaping with a linear axes resolution of 1 nm. FANUC series 30i controller is applied for the CNC. Static air bearings are selected for the movement of slides, feed screws and direct drive motors. The machine has an overall size of 1500/1380/1500 mm^3 and the stroke length of 280 × 150 mm^2 in the horizontal direction and 40 mm in the vertical direction. Surface roughness of Ra 1 nm is achieved in the turning operation on aspherical lens core of material Ni–P plate.

Moore Nanotechnology Systems [6] (USA): The company manufactures many medium size lathe machine models like Nanotech 250UPL, 350UPL and 450UPL whereas Nanotech 350FG and 500FG are 3-axis micro milling and 5-axis grinding

Fig. 1 AGIETRON MICRO NANO, nano module (Courtesy of Agiecharmilles)

Fig. 2 ROBONANO developed by Fanuc, Japan

machines respectively. 350UPL is a 4 axis lathe using oil hydrostatic slide ways. Delta Tau PC based CNC motion controller is applied. Linear feed drives use frameless, brushless DC motors having a resolution of 1 nm. Surface roughness of a cubic phase plate of Zinc sulphide material as machined on the Nanotech 350UPL is 4.112 nm Ra.

Meso and micro scale manufacturing form a middle-scale stepping stone by which the benefits of nanotechnology may be accessed. In the past 5–10 years, these meso and micro scale parts have seen increased use in medical applications, consumer products, defense applications and several other areas. A generalized approach is required for the robust miniaturization of standard size machines and manufacturing processes. Miniaturization of conventional machine tools has become a potential research area due to the high demand of the meso scale components.

1.2 Micro-manufacturing Through Micro-machines

There are many miniaturized machine tools developed so far by different research groups and institutions in different parts of the world. In Japan, micro machines are being developed as a part of a big project 'Development of a Micro-factory' [7]. Micro machines were first developed to utilize with the robotic arms in the micro factories. Micro factories are envisaged to save space and energy as they consume much less resources as compared to the existing size of machines and factories. Lu et al. [8] have developed a micro lathe turning system and tested a work material 0.3 mm in diameter that was cut to a minimum of 10 μm in diameter with a rotation speed up to 15,000 rpm. They have investigated the cutting forces and the possibility of reduction of cutting forces, thereby improving the working accuracy.

Rahman et al. [1] have assessed the machinability of micro parts by force analysis, chip analysis and tool wear criterion. They have experimented by using brass, aluminium alloy and stainless steel as a work material and carried out the experiments by varying the depth of cut, feed rate and spindle speed. One parameter was varied while the other two were kept constant in order to identify the best combination of cutting parameters. The machine tool has dimensions of 560 × 600 × 660 mm (W × D × H), and the maximum travel range is 210 × 110 × 110 mm ($X \times Y \times Z$).

Ito et al. [9] have developed a small CNC micro turning system and achieved a circularity of 50 nm and a surface roughness of 60 nm by turning a stainless steel cylinder. Kussul et al. [10] have developed micromachining centres in two stages. They call first stage to be the first sequential generation of micro equipment that is manufactured from the standard size machine tools. The second generation with smaller size than the first generation may be manufactured by using the first generation micro machines. The first generation prototype size is 130 × 160 × 85 mm^3. Some developments in the last decade in micro machines are summarized below in Table 1.

Table 1 shows the comparison of the capabilities and design aspects of existing micro machines. There are 3 micro lathes and 4 micro milling machines presented in the comparison. Out of the four milling machines, micro milling made by the AIST, Japan has less weight and size whereas Robonano machining centre has the highest resolution achieved as compared to other machines. Robonano has also achieved the highest work piece quality with a surface roughness of 1 nm. MMT from KERN, Germany has the highest spindle rotation speed of 1,60,000 rpm. In the micro lathes, MTS3 has achieved better work piece surface roughness as compared to its counterparts. However, micro lathe made by AIST, Japan is the smallest in size and the micro lathe developed by Kanazawa University, Japan, has much higher resolution than the other micro lathe machines. Figure 3 shows some machines covered in comparison.

Okazaki et al. [7] have provided a brief development history of Japanese micro factory and the benefits of utilizing it in the industry. As a first stepping stone, a micro-lathe was developed in 1996. The lathe has a size of 32 × 25 × 30.5 mm^3

Table 1 Comparison of micro machines: Features and capabilities

Machine characteristics	MTS3 (Micro Lathe)	MTS5/MTS6 (Micro milling)	Micro Lathe	Micro milling	Micro Lathe	ROBONANO α-0iB	MMT
Country of origin/make	Nano Corporation, Japan [11]	Nano Corporation, Japan [11]	AIST, Japan [7]	AIST, Japan [7]	Kanazawa University, Japan [8]	FANUC, Japan	KERN Germany
Overall size dimensions l × w × h (mm)	400 × 300 × 150 Bed size: 300 × 200	260 × 324 × 370 Bed size: 260 × 320	32 × 25 × 30.5	119 × 119 × 102	200 × 200 × 200	Length of stroke 280 × 150 × 40	2000 × 1200 × 2550 XYZ Travel (160 × 100 × 200)
Weight (kg)	25	43	0.1	0.7	10	–	500
Carried out processes	Turning of SS, boring and drilling with brass	Milling operations	Turning	Surface cutting and drilling	Facing, turning and taper turning	Milling, turning, grinding and shaping	Milling and drilling
Resolution (μm)	0.1	0.1	0.05	0.05	0.004	0.001	0.1
Maximum spindle speed (rpm)	3,000	20,000	10,000	20,000	15,000	70,000	1,60,000
Maximum work piece dimensions	Ø10 mm	50 × 50 × 16 mm^3	Ø2 mm	4 × 4 mm^2	Ø0.3 mm	–	350 × 230 mm^2 Max. workpiece weight = 30 kg
Positioning accuracy (μm)	0.5	1	0.5	0.5	–	–	± 1
Maximum work piece accuracy achieved	Circularity: 50 nm surface roughness: 60 nm	–	Surface roughness: 1.5 μm roundness: 2.5 μm	–	Surface roughness: 1 μm	Surface roughness: Ra 1 nm is achieved while turning aspherical lens core	± 2.5 μm

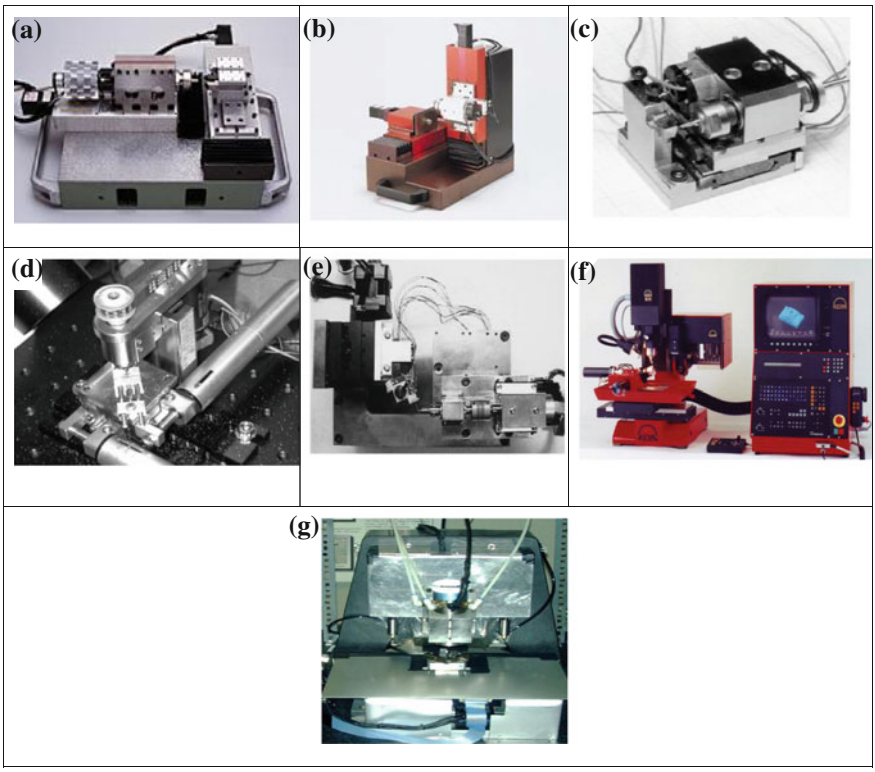


Fig. 3 **a** MTS3 [11]. **b** MTS5/MTS6 [11] (Courtesy of Nanowave, Japan). **c** AIST Micro Lathe [7]. **d** AIST Micro milling machine [7]. **e** Micro Lathe developed by Lu and Yoneyama [8]. **f** Micro machine Table top model (MMT) developed by KERN, Germany (Courtesy of KERN-microtechnic) [12], Micro milling developed in University of Manchester, UK [13]

and weighs only 0.1 kg. The machine has a feed drive resolution of 0.05 μm, positioning accuracy of 0.5 μm and can hold a maximum workpiece diameter of 2 mm. It comprises of an X–Y driving unit driven by laminated piezoactuators, a main shaft device driven by a micro motor that incorporates ball bearings with rotating accuracy of less than 1 μm. The machine has achieved a surface roughness of 1.5 μm and roundness of 2.5 μm in the workpiece turning operation of a brass rod.

Lu and Yoneyama [8] have built a micro lathe turning system of overall size $200 \times 200 \times 200$ mm^3 and weighs about 10 kg. The machine consists of X–Y and Z driving tables with an axis resolution of 4 nm. A work material of brass, Ø 0.3 mm is cut to a minimum diameter of 10 μm achieving a surface roughness under 1 μm. Kussul et al. [10] have developed micro machine tool of overall size $130 \times 160 \times 85$ mm^3 and a travel range of $20 \times 35 \times 20$ mm^3 with a resolution of 1.87 μm. Test pieces manufactured with this machine have dimensions from 50 μm to 5 mm with a tolerance range of 20 μm.

KERN [12] (Germany): KERN offers a 5-axis table top version (KERN MMT) of a large machining centre with the smallest travel range option of 160/100/200 mm^3 (X/Y/Z). Heidenhain TNC controller is applied and a feed drive resolution of 0.1 μm, positioning accuracy of ± 1 μm and the work piece accuracy of ± 2.5 μm is achieved.

NANOWAVE [11] (Japan): Many desktop models are offered by this company. MTS2, MTS3 and MTS4 are the CNC precision micro lathe systems with cross roller slide ways arranged in a 'T' configuration. MTS3 has the base size of 200 × 300 mm^2, feed drive positioning accuracy of 0.5 μm and a work piece surface roughness, Ra 0.02 μm achieved by turning brass C3604. MTS5 is a small CNC precision milling machine having a bed size of 320 × 260 mm^2. Table for each axis is supported by a set of crossed roller ways and drives through a lead screw. Machine uses a G8 controller and the positioning accuracy of feed drive is 1 μm.

2 Micro-factory as a Concept

Overall, there are plenty of benefits in miniaturization of machine tools where as there are some hidden challenges to overcome as well. To augment the desktop factory idea, large number of micro machines can be installed in a small workfloor. Manufacturing processing data and condition monitoring of micro machine tools in a micro factory are the variables of interest to run a smooth process flow. Every machine out of hundreds of micro machines will have sensing equipment and the sensors data is being compiled at one place, ideally using wireless communication systems. One or two operators can run and monitor the whole microfactory and access the machine if the fault alarms receive from any station. The control of this micro factory concept for operation, maintenance and monitoring becomes a Cyber-physical system capable of producing micro-precision products in a fully-automated manner at low cost.

2.1 Benefits of Miniaturization

1. Miniature machines will bring economical space utilization and energy saving. Individual micro machines in the Japanese micro factory take 1/50 of the space that the standard size machine tool, occupy on the shop floor. In watch manufacturing, the amount of energy consumption may be reduced to approximately 30 percent of the conventional factory by the half-miniaturization of the production systems [14].
2. Vibration amplitude will be minimized due to the reduction in mass of the moving components. Large natural frequencies will be obtained for the micro system.
3. Cutting forces will also be reduced in micro manufacturing processes that may increase the achievable accuracy of machine tool.

4. Thermal drifts that are generated by the machining process causing deformations that effect directly the accuracy of standard machines. These effects are reduced in micro machines due to the miniature nature of the components, and can often be regarded as negligible.
5. Small machines will be capable of providing high acceleration. The next generation of machine tools, will require axes to have acceleration capabilities in excess of 1 g.
6. Micro machine tool's accuracy will improve by the inherent reductions of machine component's inertia, negligible thermal drift and larger eigen-frequencies [15].
7. Consumption of raw material will be dramatically reduced in micro manufacturing. Due to the low consumption of raw material, costly materials can be utilized. Even machining of non-conventional materials like ceramics may also be possible.
8. No new research is required for the materials to be specifically used for micro manufacturing. Almost all the efforts carried out so far in micro manufacturing have made use of the same materials that are used in the macro manufacturing machines and processes.

2.2 Challenges of Miniaturization

1. The dominant hurdle for the development of micro manufacturing machines capable of machining with a very high accuracy is the identification and evaluation of the micro physical phenomena. At the micro scale, the laws of macro scale physics, no longer prevails. As different forces behave differently at the macro, micro and nano-scales, some of them are more influential at a particular scale. For example, surface forces become very important at the micro scale but their influence is negligible at the macro scale.
2. Another important issue is the assembly of the micro parts. Micro scale and multi scale products may have different challenging issues for assembling and packaging. Even the micro and meso-scale parts which will be manufactured with micro machines will have their unique requirements for fabrication and assembling. Human handling of micro parts is certainly impossible and special robotic manipulators are essentially required. All the micro machine tool developers have addressed this issue by developing micro manipulators like transfer arm and a two-fingered hand in the case of Japanese micro factory. With micro-scale components, interactive forces (e.g. van der Waals, surface tension, electrostatic) will exist between components, which instigate difficulties in manipulation and control. To overcome these problems, contact type manipulators such as ultrasonic travelling waves, or mechanical grippers; or non-contact type manipulators for example magnetic fields, aerostatic levitation, or optical trapping could be used in place of the conventional solutions [16].

3. When multi-scale parts may assemble together, the multi scale physics will play its role in design, packaging and assembly of the products. Normal design and modelling tools are not capable of handling multi scale physics and modelling. The need to interface and integrate micro scale parts with parts of a different scale may require multi-scale modelling tools to predict system-level behaviour. These implications point to a need for a departure from traditional macro scale design models and simulation tools.
4. In the micro-turning process, the rigidity or strength of the shaft decreases as the diameter of the shaft reduces. There will be a restriction of achievable minimum diameter of the shaft which can withstand the magnitude of the cutting forces for the acceptable deflection in the shaft. Lu and Yoneyama [8] have measured the deflection of work piece shaft, 50 μm in length with its reducing diameter in the micro turning operation (See Fig. 4).
5. There are many design techniques available in the literature and in practice as well for the macro scale parts. But these design techniques will not work for the micro scale parts unless the dominant physical phenomena will be addressed fully and incorporated at the design stage. Modelling tools will also be required to acquire multi scale physics. Micro manufacturing standards have not been established so far and no work is done for micro metrology and inspection of micro and meso scale parts. Being the nascent stage of this technology, there is a need to limit the gap of the state of the art micro manufacturing technology and the required knowledge. At this stage, uncertainty and risk of utilizing micro manufacturing technology for the commercial manufacturing of meso scale parts is higher. Due to the lack of modelling tools and standards for micro manufacturing and micro metrology, researchers are using their own means of modelling and simulation tools to design meso scale parts.

2.3 Micro-machine Tool Components

Following are some of the important tool components of micro machine.

Carriages and Guideways: The carriages or stages of a machine tool are components which provide movement between a probe or tool and a work piece. Carriages are railed through the guide ways in the free direction of movement. Carriages are constrained by guide ways which limit motion in any direction other than a specified linear or angular path of travel. Accurate movement of feed units is dependent on the positional accuracy determined by the encoders thereby limiting the accuracies of the traditional machine tools. Carriages are also bound to the inaccuracies present in the straightness and parallelism of the guide ways. Highly precise carriages are being designed in the industry with a positioning accuracy in nano-meters. Japanese micro machine tools have been specifically developed to be used in the microfactory concept. Linear micro stages shown in Fig. 5 were designed with piezo-actuators. Slocum et al. [17] has made a precise linear motion

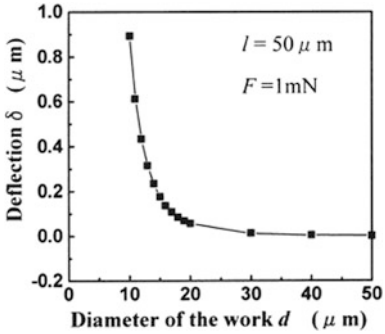


Fig. 4 Deflection estimation of work piece under cutting force [8]

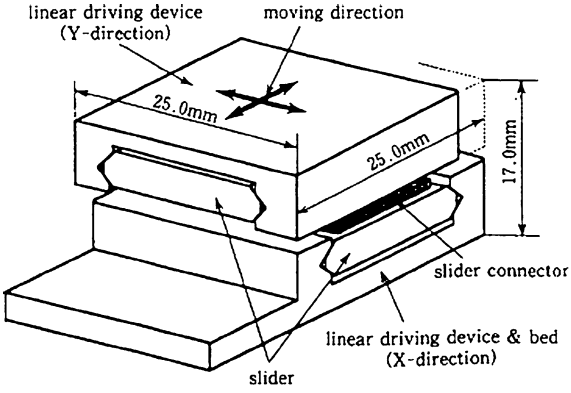


Fig. 5 X-Y Feed drives of micro machine tool [7]

carriage on which a preload is applied through the attractive force of a linear motor. The 330 mm long carriage can bear a load of more than 20 kg with a pitch error of an arc second and is supported by the six rigidly attached porous carbon air bearings (See Fig. 6). The rail is designed to be employed for standard size machine tools and macro scale applications. Mekid [18] has designed a short stroke linear high precision carriage with a 16 nm positioning accuracy. Very high axis stiffness is claimed to be achieved with a steel slide of mass 100 kg and the working motion length of 220 mm that is fully floated by three hydrostatic bearings. Yang et al. [19] have presented the design and characterization of a single-axis, low profile, piezodriven, vertical motion micro positioning stage with a travel range of 200 μm and a vertical stiffness of 6 N/μm. Mekid and Bonis [20] has presented the design of an optical delay line with a long stroke of 3 m and a 16 nm resolution. Gao et al. [21] have presented the design and characterization of a piezodriven precision micro positioning stage utilizing flexure hinges.

Bearings: A bearing is a component that allows for relative motion between two parts. There are two major classes of bearings: Contact bearings and non-contact bearings. Bearings contacts induce friction due to rolling elements, i.e., balls or rollers. Non-contact bearings are aerostatic, magnetic and hydrodynamic.

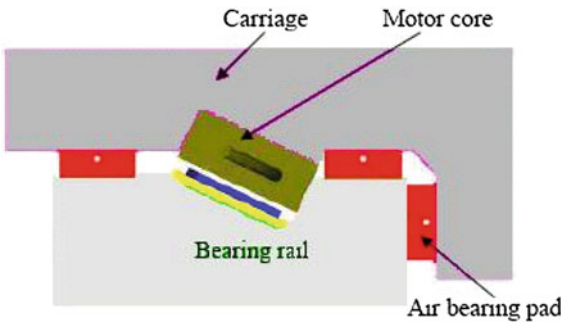


Fig. 6 Use of Aerostatic Bearings in linear motion carriage design [17]

Both classes of bearings are used in high precision machines. The industrial revolution was made possible by rotating shafts that were supported by a thin film of lubricant induced by hydrodynamic shear. Aerostatic and hydrostatic bearings rely on an external pressure source to supply gas or liquid through an inlet restrictor to a bearing pad [22]. Figure 6 shows the use of aerostatic bearing pads to bear the preload and weight of the carriage. Magnetic bearings are also been used as a non-contact type, as they have no mechanical contact with the supported component.

Spindle: Spindle unit is used to hold the tool holder in the machine tools. Spindle rotates generally up to 30,000 rpm in the modern day high speed machine tools. Influence of cutting forces is largely reduced in high speed machining. This fact may go in favour of designing ultra-precise, high speed, light weight micro spindle for machine tool. Figure 7 shows the micro spindle and the micro chuck gripping the work material 0.3 mm in diameter at a speed of 15,000 rpm.

2.4 Sensors and Actuators Used in Micro-machines and Micro-factory

Servo control should be able to handle the complete process control in a single machine. The success of machine servo control depends on the type of control applied and the interaction agility between sub-systems. Process control is extremely important for such machines to fit machine kinematics with machining process. An intelligent controller can be applied for micro-factory to control the machining process of multiple machines through CAM (Computer aided machining). CAM will include complete process planning and NC programming that can be integrated with the PC-based control system. The CAD/CAM system is directly linked to computer-aided process planning (CAPP) software tool, hence the manufacturing process can become totally automatic thus improving efficiency.

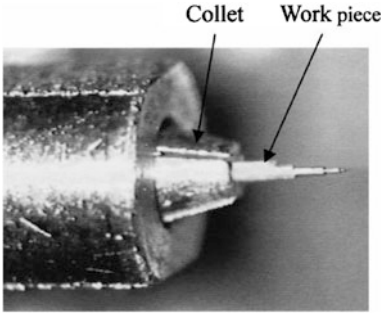


Fig. 7 Micro spindle and Chuck [8]

Modern CNC approaches employ PC-based solutions to incorporate extensive functionality in order to combine high quality and flexibility, with reduced processing time. One must also consider the processing power of the controller hardware, as too great a modularity can result in deterioration in the real-time performance of the system. Sensors and actuators employed in individual micro-machine and in the micro-factory are enlisted. This list includes the machine's inherent sensors which measure actuator's working in closed loop. These sensors and actuators are used in basic machine operation and metrology. Extra metrology sensors are also listed that will serve in micro-factory for a group of machines.

(a) **Machine operation**: (4–5 sensors per machine) Linear (bandwidth to beyond 500 kHz) or rotary encoders (High-speed rotary magnetic encoders with resolutions to 13-bit, 8192 positions per revolution) are used. Three linear encoders will work for three machine tool axis (x, y, z). One rotary encoder will measure the spindle rpm. For a three axis machine tool, these 4 sensors are enough for machine normal operation. However, the extra two rotary encoders can be employed for building up the 5-axis complete machining centre (MC). In 5-axis MC, two rotary axes are added to get maximum machining flexibility. A control model can be built for a three axis machine tool using 4 sensors. An optional study can be made using the total of six sensors for a 5 axis MC.

(b) **Machine tool Metrology**: Single machine metrology frame contains a scanning white light interferometer with CCD (Optical sensor) for calibration. Its controller sends serial data at the rate of 30 MHz for a system of resolution 6 nm and a speed of nearly 200 mm/s. Some extra positioning sensors like LVDT (Linear-voltage Differential Transformers) normally designed with sampling rate of 250 Hz output bandwidth may be employed for circularity or cylindricity measurements. Visible (normal CCD) and IR spectrum cameras for a group of machines may also be employed for operator visual aid. These cameras can be wall mounted for better visualization. Normal working of cameras is 25 frames per second.

(c) **Condition Monitoring and in-process inspection**: Capacitive micro and nano-sensing is a non-contact position measuring system and can be used for condition monitoring of machine tools especially structural and metrology frame deformation over time. Worked on an average bandwidth of up to 5 kHz and DSP 32-bit floating point, 8.3 μsec sampling rate servo controller may be implemented. As an additional condition monitoring sensor, temperature sensor like thermocouple may also employed with sampling rate of 100–1,000 Hz. In-process inspection sensors may be added in the advanced control model but not required in the basic model. In-process inspection sensors are a new concept to be used while machine operation. It can be based on laser interferometer with the sampling rate of 30 MHz

(d) **Actuators**: Some basic actuation system includes three linear motors employed for three stages. Two rotary axis may be added in case of five axis machining centre. Spindle motor needs high linearity or co-axiality and no backlash. Every stage actuator needs high stiffness in the active axis, low disturbance, availability of reverse motion and fast time response to active control. Both linear and rotary encoders explained above can be used for the actuation system. Appropriate actuation in real time is necessary. Sometimes, piezo actuators are used in micro machine tool design. Piezo-actuators work on bandwidth of up to 10 kHz.

3 Machine to Machine (M2M) Control Design in Micro-factory

An important aspect to be considered is the control system for micro manufacturing, which is increasingly being required to perform a wide variety of complicated tasks under varying operating conditions and in different environments, while at the same time achieving higher levels of precision, accuracy, repeatability and reliability [23]. In the general case of machine tools, these control techniques are commercially available into two distinct groups, i.e., open and closed architecture control. Conventional approach uses closed architecture control, however, the machine tool controller designers are turning towards more open architecture systems to address system re-configurability inventions.

The other benefit comes with multiple variety of manufacturing processes to run by the same machine. Open architecture controller may also provide the facility of programming process control at the micro factory level with re-configurable machines and CAM operations of a single machine tool simultaneously. The classical control architecture for micro machines operation in micro factory is shown in Fig. 8. However, with the advent of sensor networks and communication technologies, a micro factory CPS can be shown in Fig. 9. The disturbance shown at the input of plant can be regarded due to vibration which is compensated via active vibration control in the actuation of each axis in micro machines.

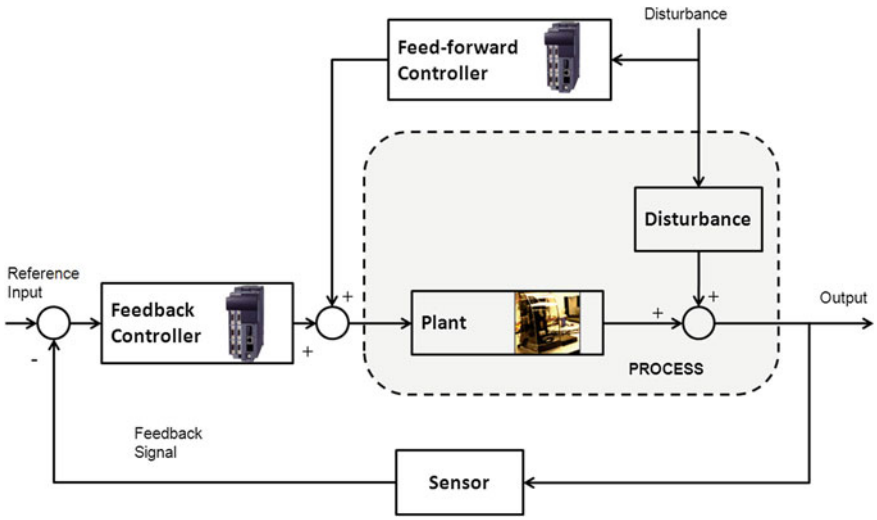


Fig. 8 Classical Control architecture of a single micro-machine

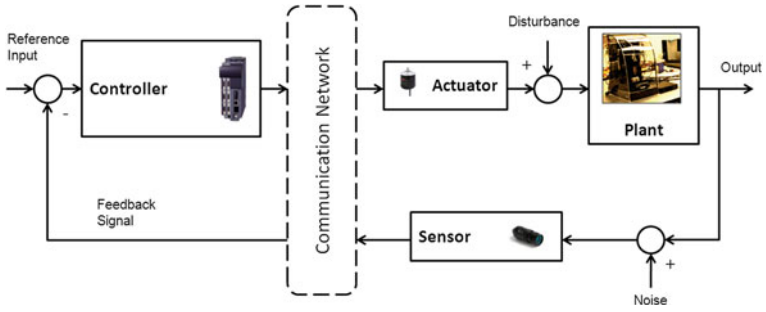


Fig. 9 Embedded Network incorporated in Control architecture

In micro factory, where multiple machines work in parallel, variety of sensors are employed for machine control, process control, metrology and calibration, condition monitoring of machine tools, assembly and integration technology at the micro-scale resulting in smooth operation of micro-factory. Single machine can be designed with a computer numerical control, but, flexible reconfigurable controllers are envisioned to control variety of processes that will lead to the development of open architecture controllers to operate micro-factory. Therefore, the control effort and algorithms have to utilize process models to improve the overall process and, ultimately, the product. Thus, we aim to introduce machine to machine (M2M) communication in the classical micro factory manufacturing test bed.

M2M communication enables micro actuator/sensor and controller devices to communicate with each other directly and without human intervention, automating management, monitoring and data collection between devices, as well as

communicating with neighboring machines. All micro-sensors communicate with a local short distance wireless network e.g. via Bluetooth piconet as well as with a centralized controller via WLAN 802.11 to exchange control/command from it.

3.1 System Controllers

In order to control the machine tool position and speed, two distinct groups of open and closed architecture exist as on today. Closed architecture controllers being more popular, have established the norm in the machine tool manufacturers market. Two popular ones are the Fanuc 15i Controller and the Mitsubishi CNC 700 series controller (See Fig. 10). Both of these controllers have high-speed reduced instruction set computing (RISC) based processors, having the ability of controlling a wide variety of systems. The CNC 700 series of Mitsubishi implements nano-control technology with RISC based processor and high-speed optical servo communication network for high-speed and high-precision real time control in 5-axis. These systems are closed architecture in nature, and the designers/manufacturers of the closed architecture controls do not anticipate moving to a more open architecture [24].

In Europe, two German companies, Bosch and Siemens are flag holders of open architecture controllers. The Bosch Rexroth IndraMotion MTX is used for controlling high-precision grinding and machining applications. While, the Siemens 840 series CNC controllers also belongs to open architecture controller group. Both of these controller designers plan to permit more advanced process control in addition to the standard servo position control. Delta Tau, D-Space, and National Instruments are some of the other open architecture systems developers actively participating in related research and development. To remain compatible with the industrial standards, the data processing for these controllers is preferably done with a combination of LabVIEW and MATLAB tools. Control of micro-machine tools requires high speed controllers, position counters having higher resolution capabilities and spline interpolation.

In order to efficiently control high bandwidth processes at micro level, increased controller speed is mandatory due to several reasons. The foremost of them is because the higher speed allows for smoother interpolation. Secondly, servoing at ultra-high resolution (on the order of 1 nm) requires that the controller is capable of tracking commands at very high rates. For example, if a machine has 1 nm resolution and is traveling at feed rate of just 1 mm/s, the controller must track an encoder pulse at a rate of 1 pulse/μs, resulting in a 1 MHz clock requirement. However, if the system is traveling at feed rate of 1 m/s, an encoder pulse is generated every ns which requires a 1 GHz clock. If this same machine has a one-meter range, the controller must be able to track a billion encoder counts, requiring a 30-bit (minimum) counter [24]. This theoretical curve for encoder speed vs. range of the encoder counts is quite significant and outside of most standard controller specifications.

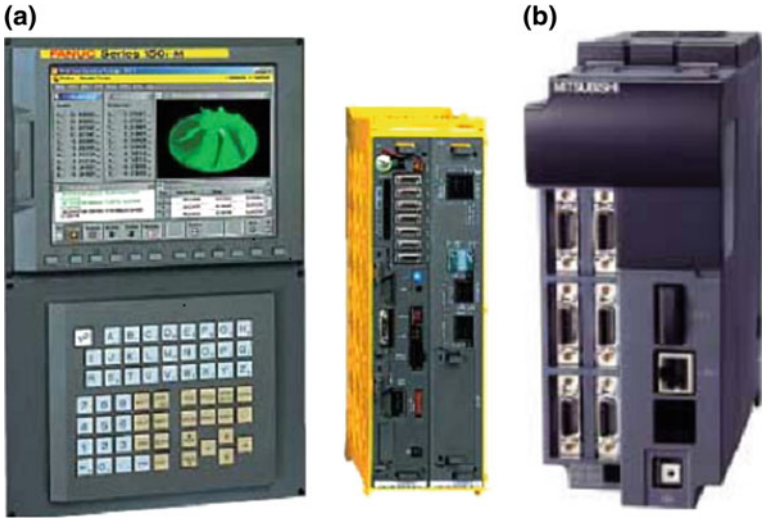


Fig. 10 **a** FANUC 15i controller (Courtesy of FANUC Inc.), **b** Mitsubishi CNC 700 controller (Courtesy of Mitsubishi Inc.) [24]

3.2 Open Architecture Controls/Control Flexibility

There is a clear dichotomy regarding the utilization of open architecture controls in the micro-manufacturing area. In application specific embedded systems and industrial solutions, closed architecture controllers are used. While in R&D centers, an open architecture and flexible platforms are specifically preferred as per research needs. PC based controllers that are operated in real-time, single-board motion controllers (SBMC) and field programmable gate arrays (FPGA) provide flexibility of processing. However, from an industry/implementation perspective, closed architecture controllers using hardware based on application specific integrated circuit (ASIC) systems are employed as the process specific hardware design has already been finalized.

There are some designs with the "closed architecture" controls such as the FANUC or Mitsubishi controllers, where a single controller can run multiple types of machines. For example, ROBOnano can be configured for milling, turning, shaping and several other operations using FANUC 15i controller. These controllers are flexible enough but with closed architecture. Some designers prefer to stick with their policies e.g. the mainstream control companies in Japan such as FANUC and Mitsubishi seems to continue producing controllers only with closed loop architecture. On the other side, German control companies such as Siemens and Bosch appear to provide a more open architecture platform but still not as open as the DSP and FPGA based controllers [24].

4 Micro-factory Cyber Physical System Architecture

Machine to machine (M2M) refers to communication technologies that allow wireless and/or wired systems to communicate with other devices of the same type [25–27]. M2M is a generalized title which does not pinpoint specific wireless or wired networking, information and communications technology. The application areas for M2M include industrial instrumentation where sensor measurements can be relayed through a network (wireless, wired or hybrid) to a supervisor application that translates the captured event into meaningful information and records in the data base. Such communication was originally accomplished by having a remote network of machines relay information back to a central hub for analysis, which would then be rerouted into a system. M2M is also visualized as a combination of automation, network and SCADA. Figure 11 shows a sensor and actuator network installed within a micro factory for control and command tracking. Position and speed sensors send information to controller which commands the actuators to nullify the error.

We are aiming to introduce machine to machine, machine to man and man to machine communication integration within micro factory CPS which will enable the flow of data between micro-machines and ultimately to micro-factory supervisor. Generally, the information flows from a machine over a network, and then through a gateway to a system where it can be reviewed and acted on. Within that basic framework, there are many different choices to make such as how a micro-machine is connected to its neighboring machine in micro-factory to perform the manufacturing tasks in a collaborative manner, what type of embedded as well as long range communication protocol is used, and how the data is interpreted. Machine to machine conversation will take most of the process planning and execution control from the human intervention. Thus making the system intelligent enough to plan and execute the tasks by itself. Human supervisor role will only be maintenance specific. Even though it can be complex, once a designer knows what it wants to do with the data (whether allow more liberty to embedded computing and decision making or to include human in the loop for critical tasks), the options for setting up the application are usually straightforward. There are four basic stages that may be shortlisted in our micro-factory CPS application. Those components are sensor data acquisition and embedded control, data transmission through a communication network, data logging and Man-In-the-loop operation.

4.1 Collection of Data and Embedded Control

The first stage for error compensation in desired manufacturing dimensions is to sense these measurements. The intelligence of microfactory CPS can be enhanced just by installing multiple sensors e.g. force sensors and accelerometers, linear and rotary encoders for position sensing, temperature sensor, visual sensing using an industrial computer system with a Modbus communication port.

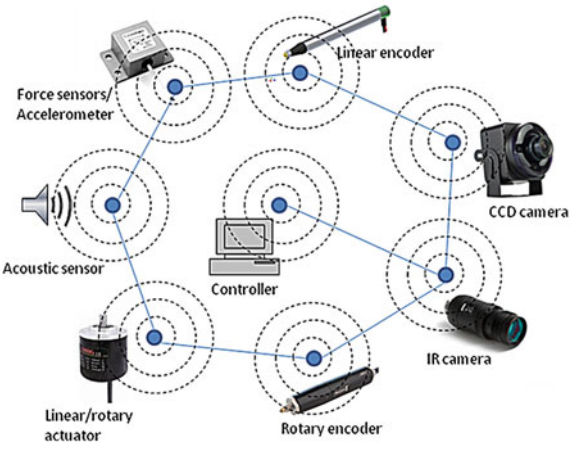


Fig. 11 Description of a stand-alone micro factory

Table 2 Comparison of WPAN and WLAN protocols

Type	Protocol	Frequency band	Data rate (Mbps)	Bandwidth efficiency (bps/Hz)	QoS	Range (m)
WPAN	Bluetooth v2.0	2.4–2.4835	2.1	2.1	No	10
	UWB	4.8–10	480	0.96	No	10
	ZigBee	2.4–2.484	250 Kbps	0.125	Yes	50
	6loWPAN	2.4–2.484	250 Kbps	0.125	Yes	50
	Wireless HART	2.4–2.4835	250 Kbps	0.125	Yes	50
WLAN	802.11a	5	6–48	2.7	Yes	70–100
	802.11b	2.4–2.484	11	0.55	Yes	35–100
	802.11g	2.4–2.484	6–54	2.7	Yes	100

The process of M2M communication in microfactory CPS begins with taking data out of a machine so that it can be analyzed and sent over a network. The controller node performs the intelligent control part for precise command following as desired. A list of WPAN and WLAN protocols is shown in Table 2. The choice of a suitable protocol depends upon the data rate, secure communication and range. Bluetooth piconet is preferred when communicating in close proximity while ZigBee or wireless HART based 802.15.4 protocol is preferred for medium range communication. In wireless HART, a time division protocol is utilized for real time communication over wireless. Also, it uses channel blacklisting to avoid interference. The availability of quality of service option can be used to give preference to certain communicating node for time/resource sharing as compared to others.

In a high end application where man in the loop supervisory control is available, it may be necessary to send a constant stream of real time data describing the machine or process. But in many cases, this is not necessary or worth the cost. In these cases, the M2M local controller should minimize the amount of data to be

sent by constantly reviewing the data, comparing it against programmable alarm limits or safety set points, and then transmitting only those real time information when a measurement is out-of-limit. For example, in micro-assembly, it is a common issue of releasing micro parts once gripped by a micro-gripper due to highly influential surface tension forces. Such information or alarm of not releasing the micro-part must reach the supervisor to go for an alternate method. In addition, the application will typically be programmed to send complete data updates on a time scheduled basis or anytime upon request from the web server for supervisory control only. This reduces the bandwidth and data rate demand for the long range communication segment.

4.2 Transmission of Data Through a Communication Network

Once the data is gathered and useful data set is picked, the next step is to transmit the data to the remote supervisor through a communication network. There are several options for transporting data from the micro-factory CPS to the network operation center. The cellular network, telephone lines, and communication satellites are all common solutions. The telephone may be the best choice if a line is already installed. Its disadvantage is usually the ongoing monthly cost and sometimes the cost and difficulty of installation beside disconnection issues. Satellite may be the most expensive solution, but is often the most reliable and the only solution for monitoring equipment remotely. The advantage of cellular networks is the ability to send large amounts of data frequently at a cheap price.

As shown in Fig. 12, a gateway is used to connect the micro-factory with the cellular or satellite network. Data security features such as authentication and access control can be managed by the gateway and the application software. The gateway can also support IP communication for an end-to-end IP network flow. Specially, when the flow of data is from supervisor to the micro-factory for supervisory control, the gateway not only converts the protocol but also transforms high-bandwidth Internet protocols to low-bandwidth wireless protocols so that the minimum data is sent over a cellular or satellite network.

4.3 Data Logging and Record Keeping

Data from the micro-factories is logged on the controller as well as to the supervisor via a long distance network. Data logging and record keeping is an essential feature of M2M communication. It is worth mentioning that the data records are matched with the historical data for performance indication and fault detection in the machine calibration. Later, these observations are used in maintenance cycle.

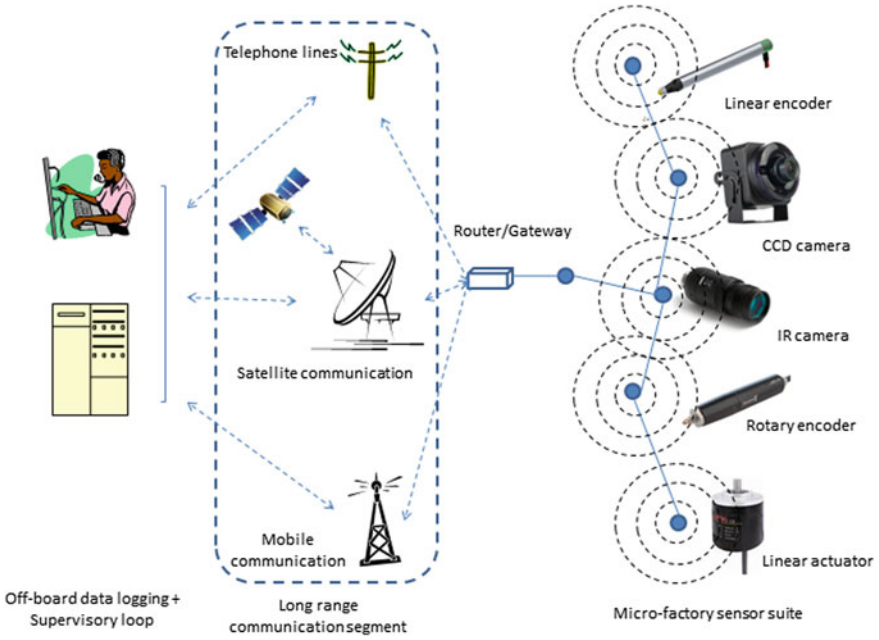


Fig. 12 Micro-factory test-bed and Supervisory control

4.4 Man-in-the-Loop Operation

In micro-factory CPS, the optimal performance can be achieved by taking into account the embedded control solution. However, if need arises, man-in-the-loop operation can be used to remotely access the system. However, despite the existence of QoS support on various long range protocols, it is very difficult to obtain real time performance. Hence, the stabilization loop in the micro-factory must incorporate embedded control loop and only few commands can be operated remotely to avoid malfunction due to time delay.

5 Design Tentative

The comparison of existing micro-machine tools shows a picture about the state of the art technology. But the designs are nascent and need considerable time in future for commercialization. Machining centers are also developed for the multipurpose applications of milling, drilling, turning and grinding. Before starting the work for the design strategy of this type of micro machining centre to be used in a micro-factory concept, familiarity with the components used in the construction and assembly of modern micro machines is discussed above.

5.1 Physical Phenomena of Scaling

Design strategy is developed with the aim of getting a very high accuracy in micro machining. The difference in this design strategy from the strategy of standard size machine lies in the introduction of second order phenomena that includes the micro physics involved between the components of the micro machine tool. As evident from the comparison of existing micro machines developed so far, the accuracies obtained from them are not improved in comparison to the standard machines. Micro machines should be developed with the aim that down scaling of machine dimensions will give more accuracies than the standard machines. But to get higher accuracies, certain design strategies have to be developed and the miniaturization issues are to be addressed. These issues deal with the magnitude of the cutting forces, thermal drift, reduced masses and inertia and down scaling of physical forces.

The common example of scaling implies that while gravitational force may be a significant phenomenon between two large bodies, surface tension forces which are not important between two large bodies become significant in the case of small/micro objects. If the scale factor *'s'* is used to describe how the physical phenomena change. All the lengths will scale by the factor *'s'*, but the volume scales differently:

$$V = L \times W \times H \tag{1}$$

$$V = S^3(L \times W \times H) \tag{2}$$

If our mechanical system scale down by a factor of 10, then volume will scale as (1/10)3, or 1/1,000. Different forces and parameters scale differently. For example, the mass scales as s^3. As mentioned, Trimmer [28] has tried to make use of the scaling of electromagnetic and electrostatic forces to build micro motors that can be used in micro machine and MEMS applications.

Micro-motors based on electrostatic force are theoretically a potential candidate, as electrostatic force scales to s0 (See Table 3). But MEMS micro motors experience demonstrate that in the micrometric range, the electrostatic motors cannot have sufficient power to be used in micro machining and especially in high speed machining. In this sense, they cannot compete with micro motors based, for example, on piezoelectric forces. However, these small motors have plenty of small applications like the microcomb drive motor [29] is being used in automobiles as an airbag sensor. The miniaturization process has some inherent problems that must be highlighted. The matter is not only size reduction but the effect of new physical phenomena on micro machining has to be identified. As mentioned, the behavior of forces changes in the micro domain as compared to the standard scale. This implies that the micro-machining cannot be handled as standard scale machining. There is a difference between the two and it needs to be verified experimentally that the theoretical downscaling results presented in Table 3 holds true. There is also a possibility of making use of these scaling results in order to enhance the micro machine's accuracy.

Table 3 Influence of physical forces at miniature level

Type of force	Formula	scaling factor	Symbols explained
Gravitational force	Gm_1m_2/r^2	s^4	G—Gravitational constant, m_i—interacting masses
Elastic force	$EA\Delta L/L$	s^2	E—Young's Modulus, A—Cross-section Area
Inertial force	$M\partial^2x/\partial t^2$	s^4	$\delta^2x/\delta t^2$—Acceleration, M—moving mass
Electrostatic force	$\varepsilon AV^2/2d^2$	s^0	ε—Permittivity, V—Voltage applied, d—Gap between electrodes
Surface tension force	γL	s^1	γ—Surface tension constant
Electromagnetic force	$BA/2\mu$	s^4, if B is constant s^2, if heat sink is used	B—Magnetic field density, μ—Permeability, A—Surface area
Thermal expansion force	$EA\Delta L(T)/L$	s^2	E—Young's Modulus, T—Temperature
Piezoelectric force	$EA\Delta L(E)/L$	s^2	E—Young's Modulus, L—Piezo length
Centrifugal force	$m\omega^2r$	s^4	ω—Constant rpm, r—Radius of gyration
Hydraulic force	ΔPA	s^2	ΔP—Pressure difference on the piston
Capillary force	$\sigma\pi d$	s^1	σ—Surface tension force, d—Capillary diameter
Van der wall force	E/r	s^{-1}	r—Atomic centre dist., E—Atom bonding energy
Viscosity force	$\frac{cA\partial x}{L\partial t}$	s^2	c—Viscosity co-efficient, L—Characteristics length
Frictional force	μmg	s^3	μ—Frictional constant, g—gravity constant

5.2 Precision Engineering Fundamentals

Machine tools performance is always measured by their capabilities in terms of accuracy, resolution and repeatability. These metrological terms are defined according to the 'International vocabulary of basic and general terms in metrology'.

Accuracy: Closeness of the agreement between the actual value resulting from an operation and a target value of the quantity. Accuracy is a qualitative description.

Uncertainty: Parameter associated with the result of an operation that characterizes the dispersion of the values that could reasonably be attributed to the quantity.

Resolution: Smallest difference between indications of displaying device that can be meaningfully distinguished.

Repeatability: Closeness of the agreement between the results of successive operations of the same quantity carried out under the same conditions.

Reproducibility: Closeness of the agreement between the results of operations of the same quantity carried out under changed conditions.

5.3 Design Strategy

The strategy started with the problem definition or the specifications of the micro machine tool. i.e., the design features and objectives. This is followed by the generation of basic conceptual ideas of the micro machine tools. The concepts are based on the precision design principles. All the initial concepts are compared according to the selection criteria. The criterion is defined on the basis of machine's static and dynamic design factors. The concept analysis will result into a final model that will be further refined and analyzed. The detailed model will be further used for the mathematical modeling of the system. The Homogenous Transformation matrices are applied to the machine tool detail design. The HTM model is then used for the optimization analysis. Objective function is defined based on the machine's relative volumetric error. From the optimization analysis, tolerances budget can be evaluated for the machine's components. The sub-systems will include machine's carriages, guide ways, bearings and spindle system etc. The modeling of micro physical phenomena or commonly called 'the second order phenomena' can also be modeled with the theoretical model of machine tool. Finite Element Analysis based static, dynamic and thermal models can be integrated with the optimization model for this purpose.

For control design, there are two different philosophies opted in precision machine design to achieve high precision as follows.

1. Design of precise mechanical structures with most of the phenomena considered as second order error sources addressed. An adjusted servo-controller will animate the system to satisfy the specifications.
2. Design of a mechanical structure with an overall satisfaction with the implementation of an expert dedicated servo controller to compensate for all errors [22].

If the robust design analysis is conducted with the complex mechanical second order phenomena addressed, the first philosophy can be applied with an adjusted servo controller. Servo control will also handle the complete process control. An intelligent controller will be applied to control the machining process through CAM (Computer aided machining). CAM will include complete process planning, and NC programming. In the case of un-modeled errors, specific controllers will be applied for an eventual compensation. Finally, the success of the machine servo control depends on the type of control applied and the interaction agility between sub-systems.

A conceptual example of a 5-axis micro milling machine with 3 translational (x, y, z) and 2 rotational stages, shown in Fig. 13. The machine comprised of a

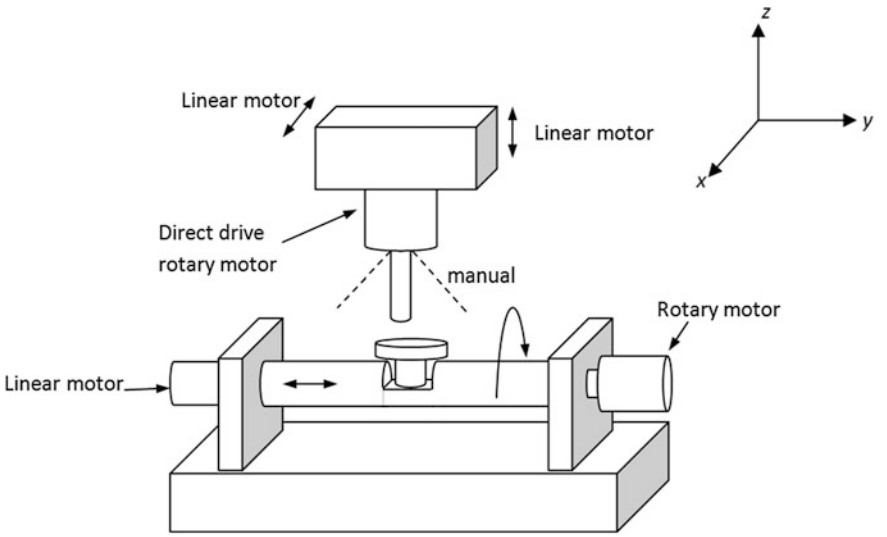


Fig. 13 5-axis micro machining center concept

central shaft that can translate in the Y axis and rotate around the same axis. This Hybrid Axis is used to mount the work piece. The axis will be mounted on air bearings on both sides for high precision.

5.4 Error Model

A robust design with an optimization method considering the volumetric error as the objective function is applied to characterize some of the key-design parameters of the machine. The method is presented with few dimensional results. A serial micro milling machine (See Table 4) with X–Y linear stages and the vertical Z slide has its overall error modeled through homogeneous transformation matrices (HTM). The motion transformation matrices from reference to work piece and from reference to the tool are shown in Eqs. (3) and (4).

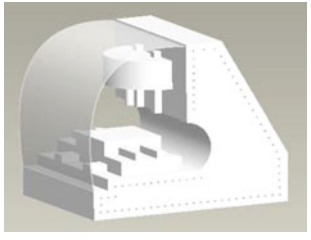
$$ {}^{R}T_{Work} = {}^{R}T_{Y}.{}^{Y}T_{X}.{}^{X}T_{Work} \tag{3} $$

$$ {}^{R}T_{Tool} = {}^{R}T_{Z}.{}^{Z}T_{Tool} \tag{4} $$

where,

$$ {}^{R}T_{Y} = \begin{bmatrix} 1 & -\varepsilon_{zy} & \varepsilon_{yy} & \delta_{xy} + a_1 + S_{xy}.Y \\ \varepsilon_{zy} & 1 & -\varepsilon_{xy} & \delta_{yy} + Y + b_1 \\ -\varepsilon_{yy} & \varepsilon_{xy} & 1 & \delta_{zy} + c_1 + S_{zy}.Y \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{5} $$

Table 4 Detailed selected concept with nominal dimensions

	Overall machine dimensions/Specs (mm)	Initial nominal dimensions (mm)
	Base size: 300 × 200 × 50 Y Carriage: $l_y \times 120 \times l_{thy}$ X Carriage: 100 × 100 × l_{thx} Work piece : 20 × 20 × 20 Column : 200 × 70 × 210 Z slide thickness: 40 Spindle: $l_s \times \emptyset 50$ Tool height : $l_t \times \emptyset 14$ Machine working volume = 150 × 80 × 20 mm^3	l_s—Spindle thickness—20 $l_{th} = l_{thx} = l_{thy}$—*X* and *Y* Feed unit thickness—20 l_y—*Y* Carriage length—250 l_d—Spindle column distance—90 l_t—Tool length—20

and a_1, b_1 and c_1 are constant offsets defined on the basis of machine design variables. δ_{xy} and δ_{zy} are straightness errors. S_{xy} and S_{zy} are the squareness angles between the respective axis and Y is the position of Y axis where these angles are amplified to yield an Abbé error in Y-direction [30].

Similar transformations are defined for Y to X and reference to Z carriages.

$$ {}^{Y}T_{X} = \begin{bmatrix} 1 & -\varepsilon_{zx} & \varepsilon_{yx} & \delta_{xx} + a_2 + X \\ \varepsilon_{zx} & 1 & -\varepsilon_{xx} & \delta_{yx} + b_2 + S_{xy}.X \\ -\varepsilon_{yx} & \varepsilon_{xx} & 1 & \delta_{zx} + c_2 + S_{zx}.X \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{6} $$

$$ {}^{R}T_{Z} = \begin{bmatrix} 1 & -\varepsilon_{zz} & \varepsilon_{yz} & \delta_{xz} + a_3 + S_{zx}.Z \\ \varepsilon_{zz} & 1 & -\varepsilon_{xz} & \delta_{yz} + b_3 + S_{zy}.Z \\ -\varepsilon_{yz} & \varepsilon_{xz} & 1 & \delta_{zz} + Z + c_3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{7} $$

For the work piece and the tool, the position vector is given as:

$$ {}^{X}T_{Work} = \begin{bmatrix} W_x \\ W_y \\ W_z \\ 1 \end{bmatrix} \tag{8} $$

$$ {}^{z}T_{Tool} = \begin{bmatrix} T_x \\ T_y \\ T_z \\ 1 \end{bmatrix} \tag{9} $$

For the ideal situation without errors, we take:

$$ {}^{R}T_{Work} = {}^{R}T_{Tool} \tag{10} $$

$$ {}^{R}T_{Y}.{}^{Y}T_{X}.{}^{X}T_{Work} = {}^{R}T_{Z}.{}^{Z}T_{Tool} \tag{11} $$

Hence,

$$ {}^{X}T_{Work} = \begin{bmatrix} W_x \\ W_y \\ W_z \\ 1 \end{bmatrix}_{ideal} = \left[{}^{R}T_{Y}.{}^{Y}T_{X}\right]^{-1}\left[{}^{R}T_{Z}.{}^{Z}T_{Tool}\right] \tag{12} $$

$$ {}^{X}T_{Work} = \begin{bmatrix} -a_1 - a_2 + a_3 - X + T_x \\ -b_1 - b_2 + b_3 - Y + T_y \\ -c_1 - c_2 + c_3 + Z + T_z \\ 1 \end{bmatrix} \tag{13} $$

By putting the ideal ${}^{X}T_{Work}$ in Eq. (3) and assuming that the work piece mounted on the x-axis stage is error free, the final vector for the tool and work piece respectively are given as follows:

$$ {}^{R}T_{Tool} = \begin{bmatrix} a_3 + \delta_{xz} + T_x - T_y\varepsilon_{zz} + T_z\varepsilon_{yz} + S_{zx}.Z \\ b_3 + \delta_{yz} + T_y + T_x\varepsilon_{zz} - T_z\varepsilon_{xz} + S_{zy}.Z \\ c_3 + Z + \delta_{zz} + T_z - T_x\varepsilon_{yz} + T_y\varepsilon_{xz} \\ 1 \end{bmatrix} \tag{14} $$

Equation (14) shows three dimensional error between global reference and tool (${}^{R}T_{Tool}$) and Eq. (15) is showing error between global reference and work piece (${}^{R}T_{Work}$). The difference of the two will give the three dimensional volumetric error.

$$ {}^{R}T_{Work} = \begin{bmatrix} a_1 + a_2 + X + \delta_{xx} + \delta_{xy} \\ + \left[(-a_1 - a_2 + a_3 - X + T_x)\left(1 - \varepsilon_{zy}\varepsilon_{zx} - \varepsilon_{yy}\varepsilon_{yx}\right)\right] \\ + \left[\left(-b_1 - b_2 + b_3 - Y + T_y\right)\left(-\varepsilon_{zx} - \varepsilon_{zy} + \varepsilon_{yy}\varepsilon_{xx}\right)\right] \\ + \left[(-c_1 - c_2 + c_3 + Z + T_z)\left(\varepsilon_{yx} + \varepsilon_{yy} + \varepsilon_{zy}\varepsilon_{xx}\right)\right] \\ - \varepsilon_{zy}\left(\delta_{yx} + b_2 + S_{xy}.X\right) + \varepsilon_{yy}(\delta_{zx} + c_2 + S_{zx}.X) \\ + S_{xy}.Y \\ \\ b_1 + b_2 + Y + \delta_{yy} + \delta_{yx} \\ + \left[(-a_1 - a_2 + a_3 + T_x - X)\left(\varepsilon_{zx} + \varepsilon_{zy} + \varepsilon_{xy}\varepsilon_{yx}\right)\right] \\ + \left[\left(-b_1 - b_2 + b_3 - Y + T_y\right)\left(1 - \varepsilon_{zy}\varepsilon_{zx} - \varepsilon_{xy}\varepsilon_{xx}\right)\right] \\ + \left[(-c_1 - c_2 + c_3 + Z + T_z)\left(\varepsilon_{zy}\varepsilon_{yx} - \varepsilon_{xx} - \varepsilon_{xy}\right)\right] \\ + \varepsilon_{zy}(a_2 + X + \delta_{xx}) - \varepsilon_{xy}(c_2 + \delta_{zx} + S_{zx}.X) \\ + S_{xy}.X \\ \\ c_1 + c_2 + \delta_{zx} + \delta_{zy} \\ + \left[(-a_1 - a_2 + a_3 - X + T_x)\left(-\varepsilon_{yx} - \varepsilon_{yy} + \varepsilon_{xy}\varepsilon_{zx}\right)\right] \\ + \left[\left(-b_1 - b_2 + b_3 - Y + T_y\right)\left(\varepsilon_{xy} + \varepsilon_{xx} + \varepsilon_{yy}\varepsilon_{zx}\right)\right] \\ + \left[(-c_1 - c_2 + c_3 + Z + T_z)\left(1 - \varepsilon_{yy}\varepsilon_{yx} - \varepsilon_{xx}\varepsilon_{xy}\right)\right] \\ - \varepsilon_{yy}(a_2 + X + \delta_{xx}) + \varepsilon_{xy}\left(\delta_{yx} + b_2 + S_{xy}.X\right) \\ + S_{zx}.X + S_{zy}.Y \end{bmatrix} \tag{15} $$

The total error vector is found in Eq. (16) and the individual equations of its components P_x, P_y and P_z are given by Eqs. (17–19) (Fig. 14).

$$P_e = \begin{bmatrix} P_x \\ P_y \\ P_z \end{bmatrix} = {}^{R}T_{Tool} - {}^{R}T_{Work} \tag{16}$$

$$\begin{aligned} P_x = & -a_1 - a_2 + a_3 + T_x - X - \delta_{xy} - \delta_{xx} + \delta_{xz} \\ & - (c_2 + \delta_{zx} + S_{zx}.X)\varepsilon_{yy} + T_z\varepsilon_{yz} + \left(b_2 + \delta_{yx} + S_{xy}.X\right)\varepsilon_{zy} \\ & - (-c_1 - c_2 + c_3 + T_z + Z)\left(\varepsilon_{yy} + \varepsilon_{yx} + \varepsilon_{xx}\varepsilon_{zy}\right) \\ & - \left(-b_1 - b_2 + b_3 + T_y - Y\right)\left(\varepsilon_{xx}\varepsilon_{yy} - \varepsilon_{zy} - \varepsilon_{zx}\right) \\ & - T_y\varepsilon_{zz} - (-a_1 - a_2 + a_3 + T_x - X)\left(1 - \varepsilon_{yy}\varepsilon_{yx} - \varepsilon_{zy}\varepsilon_{zx}\right) \\ & + S_{zx}.Z - S_{xy}.Y \end{aligned} \tag{17}$$

$$\begin{aligned} P_y = & -b_1 - b_2 + b_3 + T_y - Y - \delta_{yy} - \delta_{yx} + \delta_{yz} \\ & + (\delta_{zx} + c_2 + S_{zx}.X)\varepsilon_{xy} - T_z\varepsilon_{xz} - (a_2 + X + \delta_{xx})\varepsilon_{zy} \\ & - (-c_1 - c_2 + c_3 + T_z + Z)\left(-\varepsilon_{xy} - \varepsilon_{xx} + \varepsilon_{yx}\varepsilon_{zy}\right) \\ & - (-a_1 - a_2 + a_3 + T_x - X)\left(\varepsilon_{xy}\varepsilon_{yx} + \varepsilon_{zy} + \varepsilon_{zx}\right) \\ & + T_x\varepsilon_{zz} - \left(-b_1 - b_2 + b_3 + T_y - Y\right)\left(1 - \varepsilon_{xy}\varepsilon_{xx} - \varepsilon_{zy}\varepsilon_{zx}\right) \\ & + S_{zy}.Z - S_{xy}.X \end{aligned} \tag{18}$$

$$\begin{aligned} P_z = & -c_1 - c_2 + c_3 + T_z + Z - \delta_{zy} - \delta_{zx} + \delta_{zz} \\ & - \varepsilon_{xy}\left(\delta_{yx} + b_2 + S_{xy}.X\right) + T_y\varepsilon_{xz} + (a_2 + X + \delta_{xx})\varepsilon_{yy} \\ & - (-c_1 - c_2 + c_3 + T_z + Z)\left(1 - \varepsilon_{xy}\varepsilon_{xx} - \varepsilon_{yy}\varepsilon_{yx}\right) - T_x\varepsilon_{yz} \\ & - (-a_1 - a_2 + a_3 + T_x - X)\left(-\varepsilon_{yy} - \varepsilon_{yx} + \varepsilon_{xy}\varepsilon_{zx}\right) \\ & - \left(-b_1 - b_2 + b_3 + T_y - Y\right)\left(\varepsilon_{xy} + \varepsilon_{xx} + \varepsilon_{yy}\varepsilon_{zx}\right) \\ & - S_{zx}.X - S_{zy}.Y \end{aligned} \tag{19}$$

6 Robust Design Optimization

Robust Design is defined as a process of making a product insensitive to the effects of variability without actually removing the sources of disturbance. Robust design analysis has to reach a trade-off point to ensure the best compromise between different conflicting objectives. Optimization is defined as the setting of design variables such that the design is least sensitive to the effects of external variations or noise [31]. Optimization at the early design stage gives the designer a precise prediction of the outcome of the design process. A general optimization problem can be written as follows:

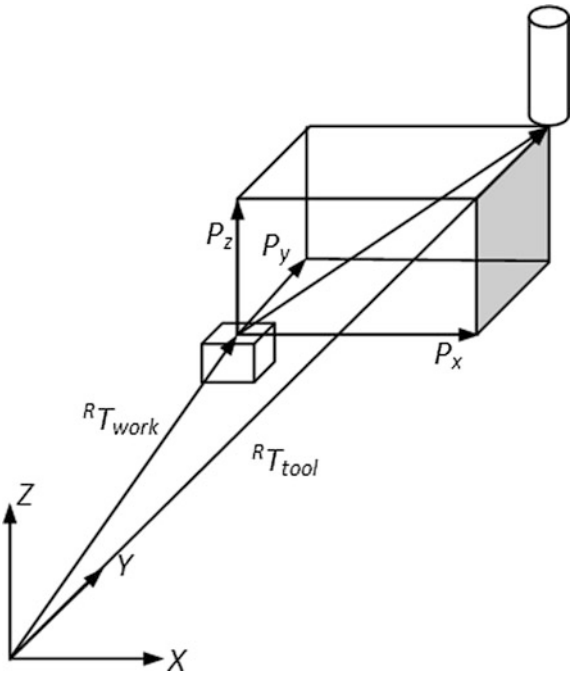


Fig. 14 Description of volumetric error

$$\text{subject to} \begin{cases} \min f_i(x) \\ g_i(x) \geq 0 \\ x_{iU} \geq x_i \geq x_{iL} \end{cases} \tag{20}$$

where, x_i is the vector of variables which may be continuous, discrete or integer with upper and lower limits, $f_i(x)$ are objectives and $g_i(x)$ are the inequality constraints. Design of Experiments (DOE) are defined as a process for generating data that utilizes a mathematically derived matrix to methodically gather and evaluate the effect of numerous parameters on a response variable [31]. Mishima [32] have modeled a miniaturized lathe through HTM and performed a tolerance analysis with the help of Taguchi Design of experiments technique.

Optimization algorithms treat the initial design of experiments data and start the search for the optimum point. 'Multi-Objective Genetic Algorithm' (MOGA) is selected as the optimization Algorithm. It has the ability of multi-objective search with fast convergence and continuous objective function penalization. Any type of variables including continuous, discrete or integers can be used. In this study, initial data is tabulated using different DOE techniques such as Taguchi, Random sequence and reduced factorial. These techniques are employed with the Multi-objective Genetic Algorithm to find out the optimum design variables for micro milling machine.

Various mechanical concepts were considered according to the functional requirements of the proposed micro machine. The concept presented in Table 4

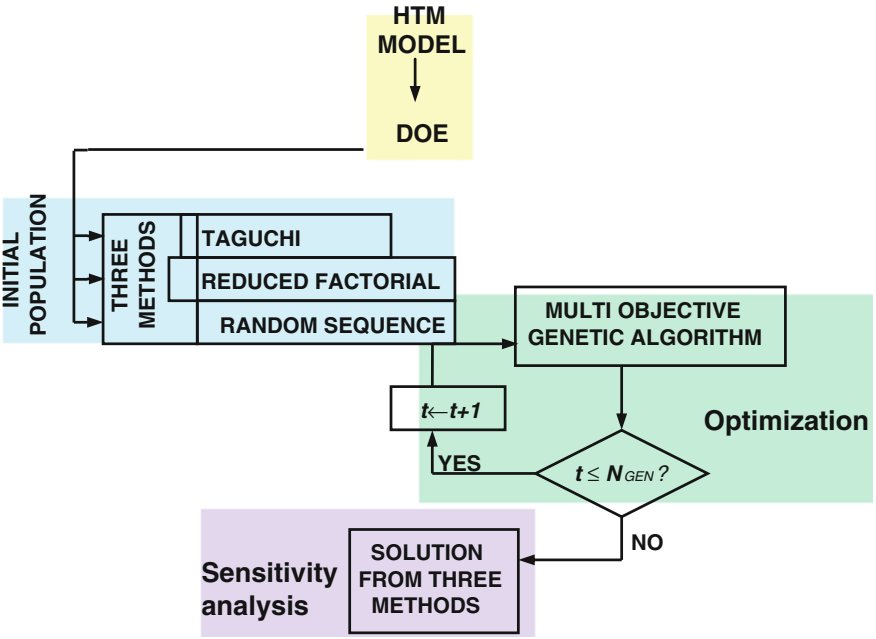


Fig. 15 Multi_objective Genetic Algorithm based optimization algorithm

was selected due to the simplicity in design and static and dynamic specifications. Dynamic characteristics become important when the tool is in contact with work piece generating cutting forces. The overall geometric errors of this concept are modeled using Homogenous Transformation Matrices (HTM) and the objective function is defined as:

$$\delta_{error} = \left(P_x^2 + P_y^2 + P_z^2\right)^{1/2} \tag{21}$$

where Px, Py and Pz are the overall volumetric errors in x, y and z directions respectively defined in Eqs. (17–19). There are five design variables chosen as an example, for the system of interest (See Table 4), three axis positions and 21 geometric errors associated with three axis of machine tool. The five design variables are evolved from the offset lengths a1, b1, c1, a2, b2, c2, a3, b3, c3, Tx, Ty and Tz. Nominal values of design variables are searched using pre-selected ranges in the optimization process (See Table 4). The procedure of calculation is shown in Fig. 15. The sensitivity analysis is carried out using the results found in the case of Taguchi method. The sensitivity of each design variable on the objective function is shown in Fig. 17; each variable is varied over a range while the other ones are kept fixed. The tolerance could be drawn from those ranges after a careful analysis. Objective function defined in Eq. (21) is used by the optimizer to minimize it for error convergence, i.e.,

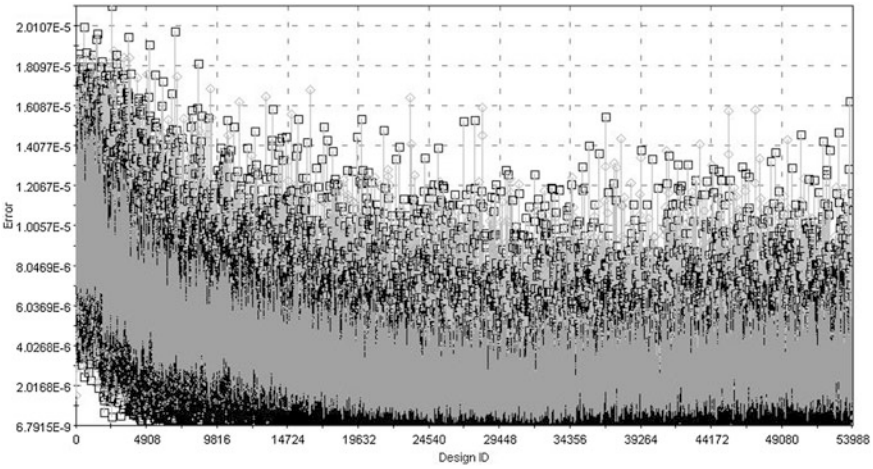


Fig. 16 History chart

Table 5 Three results with three different DOE techniques

Design variables	Initial values (mm)	Pre-selected range set	Convergence step		
			Taguchi	Reduced factorial	Random sequence
l_s	20	18–30	19	18	27.9
l_{th}	20	17–25	18	21	18
l_y	250	220–280	251	227	254
l_d	90	75–105	105	75	98
l_t	20	15–25	19	20	16
δ_{error}	–	–	6.79 nm	14.63 nm	6.93 nm

$$\delta_{error} \rightarrow 0 \tag{22}$$

Three sets of solutions are obtained by the three methods depicted in Fig. 15. An analysis of the results according to requirements will help in choosing the most adequate solution. All the three solutions have reached almost the same minimum objective value for their respective particular position (X, Y, Z) belonging to the machine workspace. Multi Objective Genetic Algorithm is used to run the model as optimizer. It is selected for its ability of fast convergence, directional cross-over and objective function penalization. Figure 16 shows the optimizer convergence towards the minimum objective function and the history of the designs.

Figure 18 shows the values of the design parameters after the convergence solution step is completed. Output parameters P_x, P_y and P_z are constrained to 1 μm individually. Designs with higher value than 1 μm of any one of them result in violation of the objective function. Data in Fig. 17 highlight different color coding for the invalid designs.

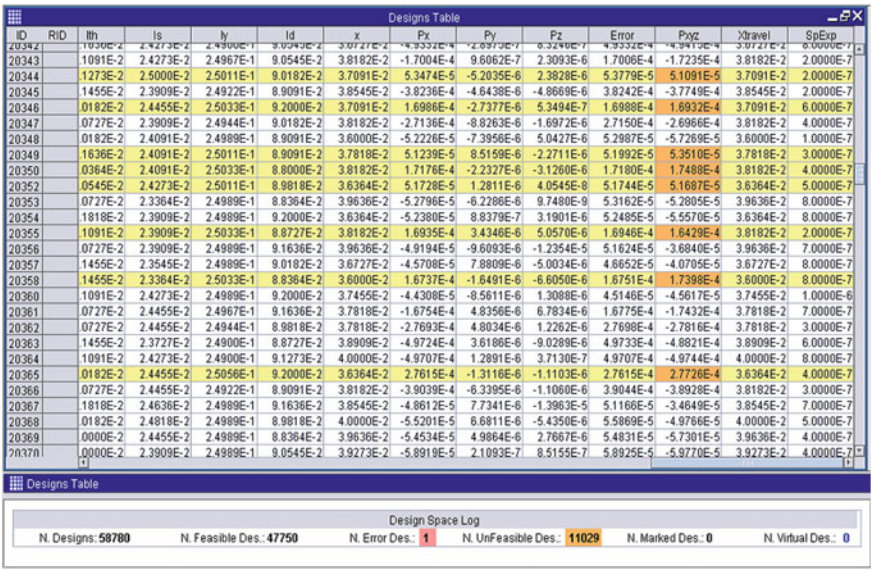


Fig. 17 Example of color coding for feasible and unfeasible designs

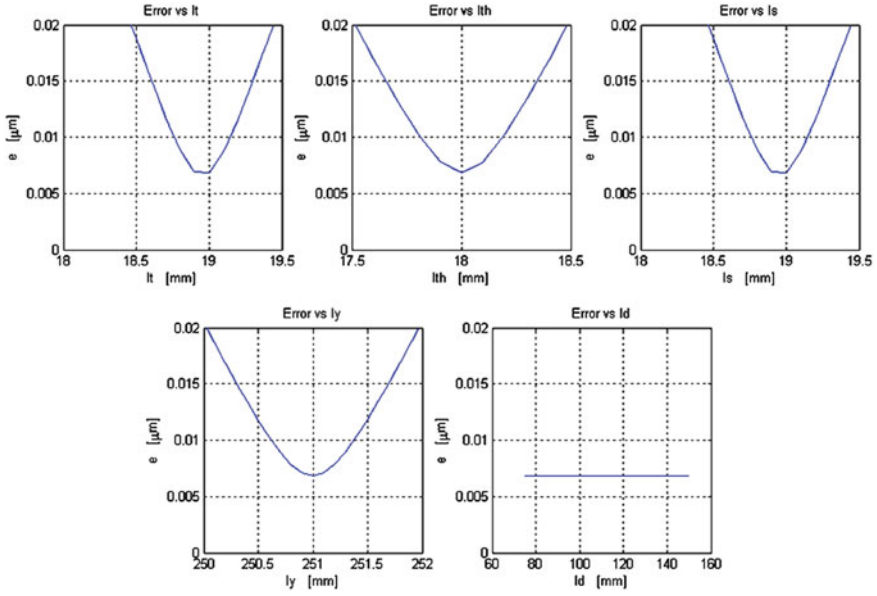


Fig. 18 Sensitivity analysis of design variables (Volumetric error vs. design parameter tolerance range)

The parameters 'l_t', 'l_{th}' and 'l_s' are found to be more sensitive design parameters than the others as shown in Fig. 18. Their tolerance band is about 1 mm for the maximum objective value of 0.02 μm. 'l_d' has been an insensitive parameter in this first analysis, as the spindle-column distance of machine tool has a specific stiffness, e.g. cantilever and will only be fully designed with dynamic analysis. Table 5 is showing the optimized design parameters using three different design of experiments techniques and the calculated objection function value in each case.

7 Conclusion

Development of high precision micro-machines is an important area where bulk micro parts can be manufactured at a high rate. Many small micro-machines that can be installed in a micro-factory environment cannot work in isolation and all the re-configurable micro-machines needs certain intelligence to work with flexibility and in an optimized fashion. For that matter, M2M communication is envisioned to run the smart micro-factory and selected parameter's data flow for the supervisory control and reporting. M2M communication is envisaged with wireless Bluetooth technology having appropriate data rates to accommodate the required operational and condition monitoring data of micro machines. A cellular network as a gateway is suggested to be used to connect the micro-factory with the remote supervisor. Data security features such as authentication and access control can be managed by the gateway through IP communication for an end-to-end IP network flow. This work also presents a robust design method of an example micro-machine tool in a micro-factory CPS using volumetric error optimization. This volumetric error is found to be critical for getting high precision to be embedded as an error compensation tool. Furthermore, three methods were used to generate the initial population of weak non-linear equations formed from the HTM model. The sensitivity investigation of the design variables of a three axis micro-machine tool reveals respective tolerance bands for the design parameters.

References

1. M.A. Rahman, M. Rahman, A.S. Kumar, H.S. Lim, CNC microturning: an application to miniaturization. Int. J. Mach. Tools Manuf. **45**(6), 631–639 (2005)
2. Y. Takeuchi, H. Yonekura, K. Sawada, Creation of 3-D tiny statue by 5-axis control ultra-precision machining. Comput. Aided Des. **35**(4), 403–409 (2003)
3. http://www.agie.com. Accessed July 2013
4. http://www.primacon.de. Accessed July 2013
5. http://www.fanuc.co.jp. Accessed July 2013
6. http://www.nanotechsys.com. Accessed July 2013

7. Y. Okazaki, N. Mishima, K. Ashida, Microfactory-concept, history and developments. J. Manuf. Sci. Eng. **126**(4), 837–844 (2004)
8. Z. Lu, T. Yoneyama, Micro cutting in the micro lathe turning system. Int. J. Mach. Tools Manuf. **39**(7), 1171–1183 (1999)
9. S. Ito, D. lijima, A. Hayashi, Precision turning on a desk - Micro turning system. Paper presented at the 4th Euspen international conference, Glasgow, 2004
10. E. Kussul, T. Baidyk, L. Ruiz-Huerta, A. Caballero-Ruiz, G. Velasco, L. Kasatkina, Development of micromachine tool prototypes for microfactories. J. Micromech. Microeng. **12**(6), 795–812 (2002)
11. http://www.nanowave.co.jp. Accessed July 2013
12. http://www.kern-microtechnic.com. Accessed July 2013
13. T.I. Ogedengbe, A contribution to the design and operation of a micro milling machine. PhD thesis, School of Mechanical, Aerospace and Civil Engineering, University of Manchester, UK, 2010
14. N. Kawahara, T. Suto, T. Hirano, Y. Ishikawa, T. Kitahara, N. Ooyama, T. Ataka, Microfactories; new applications of micromachine technology to the manufacture of small products. Microsyst. Technol. **3**(2), 37–41 (1997)
15. S. Mekid, A. Gordon P. Nicholson, Challenges and rationale in the design of miniaturised machine tool. Paper presented at the international MATADOR conference, UMIST, Manchester, 2004
16. L. Alting, F. Kimura, H.N. Hansen, G. Bissacco, Micro engineering. Ann CIRP **52**(2), 635–657 (2003)
17. A. Slocum, M. Basaran, R. Cortesi, A.J. Hart, Linear motion carriage with aerostatic bearings preloaded by inclined iron core linear electric motor. Precis. Eng. **27**(4), 382–394 (2003)
18. S. Mekid, High precision linear slide. part 1: design and construction. Int. J. Mach. Tools Manuf. 40(7), 1039–1050 (2000)
19. R. Yang, M. Jouaneh, R. Schweizer, Design and characterization of a low-profile micropositioning stage. Precis. Eng. **18**(1), 20–29 (1996)
20. S. Mekid, M. Bonis, Conceptual design and study of high precision translational stages: application to an optical delay line. Precis. Eng. **21**(1), 29–35 (1997)
21. P. Gao, S.M. Swei, Z. Yuan, A new piezodriven precision micropositioning stage utilizing flexure hinges. Nanotechnology **10**(4), 394–398 (1999)
22. A. Slocum, *Precision Machine Design* (Prentice Hall, Englewood Hill, 1992)
23. M.H. Smith, A.M. Annaswamy, A.H. Slocum, Adaptive control strategies for a precision machine tool axis. Precis. Eng. **17**(03), 192–206 (1995)
24. K.F. Ehmann, D. Bourell, M.L. Culpepper, T.J. Hodgson, T.R. Kurfess, M. Madou, K. Rajurkar R.E. Devor, *WTEC Panel Report on International Assessment of Research and Development in Micromanufacturing* (World Technology Evaluation Center (WTEC), Inc, Baltimore, 2005)
25. Z.H. Khan, J.M. Thiriet, D.G. Catalot, Wireless Network architecture for Diagnosis and Monitoring Applications. Paper presented at 6th IEEE Consumer Communications and Networking Conference (CCNC), Las Vegas, USA, 2009
26. S. Dye, Machine-to-Machine (M2M) Communications (2013), http://www.mobilein.com/M2M.htm. Accessed 2013
27. H. Chao, Y. Chen, J. Wu, Power saving for machine to machine communications in cellular networks. Paper presented at the IEEE GLOBECOM Workshops (GC Wkshps), 2011
28. W.S.N. Trimmer, Microrobots and micromechanical systems. Sens. Actuators **19**(3), 267–287 (1989)
29. W.C. Tang, T.C.H. Nguyen, R.T. Howe, Laterally driven polysilicon resonant microstructures. Sens. Actuators **20**(1–2), 25–32 (1989)

30. A.C. Okafor, Y.M. Ertekin, Derivation of machine tool error models and error compensation procedure for three axes vertical machining center using rigid body kinematics. Int. J. Mach. Tools Manuf **40**(8), 1199–1213 (2000)
31. W.Y. Fowlkes, C.M. Creveling, *Engineering methods for robust product design using taguchi methods in technology and product development* (Addison-Wesley Publishing Company, Boston, 1995)
32. Mishima N, Robustness evaluation of a miniaturized machine tool. Paper presented at the DETC, 25th design automation conference, Las Vegas, 1999

Part II
Health Care and Medicine

The Development of Cyber-Physical System in Health Care Industry

Arni Ariani and Soegijardjo Soegijoko

Abstract A cyber physical system involves the combination of sensors, actuators, and computation modules to solve issues that lie across the physical and computational areas. This emerging technology will lead to a significant improvement in the health care industry and also will enhance the quality of life of our communities, including older and disabled persons. Our objective in this chapter is to give insights from the current research to provide future perspectives for scientific research and development. We reviewed the current research and inventions in the field of cyber physical systems (CPS) focusing on the health care industry where computational intelligence is used for decision support. In this chapter, we discuss the current state of the art and trends in cyber physical system in health care industry and summarize the issues that need to be overcome. We conclude by identifying the future challenges in this technology that needs to be addressed in order to identify and facilitate priority research in this emerging field.

Keywords Cyber physical system · Telehealthcare · Wireless sensor networks

A. Ariani (✉)
Indonesian Biomedical Engineering Society (IBES), Biomedical Engineering Program, School of Electrical Engineering and Informatics, Institut Teknologi Bandung (ITB), Jalan Ganesha 10, Bandung 40132, Indonesia
e-mail: arniariani@yahoo.com

S. Soegijoko
Biomedical Engineering Program, School of Electrical Engineering and Informatics, Institut Teknologi Bandung (ITB), Jalan Ganesha 10, Bandung 40132, Indonesia
e-mail: soegi@ieee.org

Z. H. Khan et al. (eds.), *Computational Intelligence for Decision Support in Cyber-Physical Systems*, Studies in Computational Intelligence 540,
DOI: 10.1007/978-981-4585-36-1_3,

1 Introduction

A new research field has found that the cyber-physical system (CPS) has connected the physical and cyber worlds into one closed loop system [1]. CPS has a contribution to empower the physical world with the capabilities to perform self-organizing tasks and remote control operations across the network [1]. It comprises of environmental aware, embedded computing, networks communication, networks controlling and other systems engineering [1].

CPS applications can be found in different area of research including: high confidence medical devices and systems, traffic control and safety, advanced automotive systems, process control, energy conservation, environmental control, avionics, instrumentation, critical infrastructure control, distributed robotics (telepresence, telemedicine), defense systems, manufacturing, and smart structures [2].

With recent significant development in information and communication technology, further CPS applications have also recently being developed, such as: e-health and telehealth devices and systems, various types of mobile e-health systems, as well as ubiquitous health (u-health) systems.

In this book chapter, the authors review the implementation of CPS in the field of healthcare and list the pros and cons of different approaches. Moreover, the authors also summarize the overcoming challenges and provide the solutions for further work.

2 Background

Currently, the world is facing the most serious concerns in healthcare including the rise of healthcare costs, the growth of aging population and the emergence of chronic diseases [3].

In 2050, the percentage of Australians aged 65 years and older is projected to increase to 22.7 % of the forecast population of 35.9 million people from 13.5 % of 22.2 million people in 2010. This trend indicates that there will be an overwhelming demand on healthcare services for dependent older people [4]. A continuing focus on these issues will ultimately support the research and development of novel technologies that may become a solution to an unsustainable healthcare model [3].

Most companies in the field of healthcare are looking for innovative techniques that would deliver quality healthcare service, on time, with a cost-effective approach in the most efficient way [3]. The opening of new opportunities in healthcare field is being driven by recent advancements in wireless communications [5], mobile computing [6] and sensing technologies [7, 8]. The target goals are to explore the capabilities of different types of wireless technologies and solve the complexity of both wireless and mobile networks for various applications in healthcare industry, especially in relation to prevention, medical diagnosis, treatment, and patient care [9].

Recently, the wireless sensor network (WSN) has emerged as one of the important aspect in healthcare field. It has some advantages, including the following:

- It provides greater flexibility for both medical practitioner and patients since the medical practitioner can remotely monitor the physiological changes without disrupting the daily routine of the patients [10].
- It offers significant economical benefits and allows better management of equipment assets to support both of the medical practitioners and patients [11], since, it has the ability of identifying the early signs of diseases, monitoring the evolution of diseases, managing the rehabilitation process and ensuring the drug therapy recommendation achieves the desired effects [12].
- It supports a paperless environment by replacing the usage of paper and helps to reduce the possibility of errors caused by manual operation by allowing the seamless integration of physical data with context aware information system [13].

3 Personal Emergency Response System

Medical alarm or medical emergency response system is most familiar, known as personal emergency response system (PERS), allows an individual to seek help for emergencies [14]. Several studies have revealed that the availability of PERS could increase the feeling of safety, which enables older people to remain in their own home and effectively decrease the cost of healthcare services [15, 16].

A study conducted by Bernstein et al. [17] found that the usage of PERS system among older people has dropped the death rates by 34–12 % (since the patient can get the medical treatment within 60 min after having an accident), shorten the length of hospital stay by more than 59 % (14.4 vs. 5.9 days for 1 year before and after using the PERS device), produced $7.19 of direct savings in healthcare costs for each dollar spent on PERS.

Mann et al. [15] revealed that PERS was mostly used by women (70.7 % out of 93 participants) compared with men (29.3 % out of 93 participants). All the recruited users were aged over 60 years (mean: 79.3 years, standard deviation: 9.3 years). In addition, the study also reported that this PERS device has increased the feeling of safety (59 out of 78 participants, 75.6 %) and reduced the worries among family members (54 out of 78 participants, 69.2 %).

3.1 Current Technologies in PERS

There are three main components of traditional PERS [14, 18]: radio transmitters on a small-sized, a telephone is connected to the console, a call center for emergency response (for example: an emergency dispatch center-called public safety answering point [PSAP]).

The details on how the system works are presented below [14]:

- The traditional PERS is embedded into wristwatch or placed into an enclosure.
- In the case of emergency situation, the button on the PERS device is pressed to send a signal to the console.
- Once the signals have been received, the console will automatically make a phone call for emergency assistance.
- Finally, the operator staff will examine each case on an individual's situation and medical history to determine whether the emergency can be handled over phone alone or the notification must be sent to both ambulance crews and medical staffs.

The drawback of the traditional PERS is manual activation of emergency notification in the case of accident. This means that there is a possibility of failure to recognize that the accident has occurred when the user is unable to press the help button due to severe injuries, loss of consciousness or simply forgot to wear the call button or if the accident occurs outside the house [14, 19].

Unlike the traditional one that is programmed to dial the PERS service center only, the newer PERS is able to coordinate calls to preferred phone numbers when the emergency alert is received. Clearly this feature gives more safety since it can forward the call to the next emergency contact when the first name on the list is busy or is unable to respond the call [14].

The main disadvantage of commercial PERS device is that it allows two-way voice communications between the user and the call center via console unit over very short distances [14]. This may lead to annoying situations when the operator staff could not hear the voice when the user is far away from the console unit. Few PERS vendors (Philips Lifeline, AlertOne) prepared to address this issue by upgrading the existing PERS system with voice extender module. This module has the same function as console unit and can be installed in different rooms around the home.

Recently other vendors, for instance LogicMark and Medical Alarm Concepts, have made another contribution by embedding a cordless speakerphone into a pendant [14]. This approach enhances voice communications over a greater range. The user can still communicate with their family or call center staff even though they are 182 m away from where the console unit is placed. However, additional problems exist for such pendant, such as: it needs larger size and weight, with lower battery lifetime and less water resistance.

Hamill et al. [19] conducted a study that investigates the feasibility of using an automated, hands-free, dialogue-based PERS prototype for improving the usability and efficiency of PERS system in the event of an emergency. The proposed system incorporates a combination of ceiling mounted microphone array, an open-source automatic speech recognition engine, and an automated dialog with yes or no response. The content of dialog follows the Life-line Systems Canada call center protocol. The answer for each question will determine what actions need to be taken to respond to user's condition. The following is the example of question in

the proposed system: Hello. This is your automated monitoring system. Do you need help? Please say yes or no.

Three stages of testing are involved in this research. First, the performance of the system is tested with either single-or multi-microphones. A better accuracy of 49.9 % (922 out of 1,846 words) is achieved when using multiple microphone with beam-forming. Next, the accuracy of word detection is tested in two settings with voice interference from a bubble kettle and a setting with free noise interference. In this test, four males and five females aged between 20 and 30 years were asked to reply with "yes" or "no" to the question. The overall result revealed that the rate of accuracy increased to 93 % (98 out of 108 words) when combining the beam-former with a close-ended questions. Finally, the efficacy of the prototype dialog is tested in three different scenarios: an accident involved an injury that required an ambulance to be called, a fall event but prefers to contact their daughter, a false notification. Samples are taken from four healthy adults aged between 20 and 30 years.

While the recognition rate was above 78 %, their subject sample size was far too small and only involved healthy adults to give reliable results. But still, this research provides a possibility of implementing automated PERS system to assist older people during the emergency situation in a real-world environment.

3.2 Discussion

Mann et al. [15] identified three main barriers on pursuing older people to use PERS in their daily life: lack of perceived need, cost and lack of device knowledge (as shown in Table 1).

3.2.1 Perception of Need

The perception of need is a key to convince older people to use PERS in their daily life [14]. An interesting presentation about the benefits of PERS in the case of emergencies (such as fires, gas leaks, and burglaries) can make significant improvements in the decisions older people make, especially if the technology offers an opportunity to become older and independent in their own home.

3.2.2 Cost

The penetration of PERS on the market is affected by funding availability and the consumer's income level [20]. The whole costs must be reasonable when comparing to the expenses that must be paid for treatment and recovery from injuries caused by accidents.

Table 1 Barriers on using PERS from the perspective of older people [15]

	PERS users (n = 63)	PERS nonusers (n = 409)
Lack of perceived need	35 (55.6 %)	233 (57.0 %)
Cost	26 (29.9 %)	151 (37.0 %)
Lack of knowledge of device	7 (8 %)	97 (23.7 %)

According to the report by Pew et al. [21], financial support schemes can be divided into four categories:

- The grants and contracts are directly provided by the government.
- The cost is reimbursed through a national health insurance scheme (i.e. Medicare in United States).
- The cost is covered by private health insurance.
- All cost is paid by consumer.

The installation and monthly fee for PERS may vary for different countries, for example: it is fully covered by Malta government while in United States the user has to allocate at least $500 for installation fee and $45 for monthly payment.

3.2.3 A Better Understanding of the Device

A simple guideline that covers all aspect of PERS (including the installation process and how it works) might affect the decision of older people to continuously use it on daily basis [14, 15].

4 Automated Fall Detection

A fall can be defined as an unanticipated event causing an individual to lie down on the lower surface, floor or ground [22]. For older people, falling can results in devastating consequences both physically and emotionally.

The 2009–2010 financial year data in Australia reported that falls contribute significantly to injury hospitalizations (38 %, n = 161,147) and occur more frequently with increasing age [23]. Fifty-two percent (83,768 out of 161,147 people) of the documented falls are experienced by older people aged 65 years and over. Women (69.4 % out of 83,768) are more likely than men (30.1 % out of 83,768) to experience fall-related injuries [23]. The most commonly reported injuries due to falls are fractures (n = 49,325) [23]. Fracture of femur accounted for the highest incidence of admissions (n = 19,429). Fracture of forearm had the second-highest number (n = 6,513) and closely followed by fracture of lumbar spine and pelvis (n = 6,513). In addition, the report also revealed that the percentage of indoor falls (67.4 % out of 22,744) was reported to higher than those taking place outdoors [23]. The most common for these indoor falls is in the home (49.3 %) and

followed by falling in residential institution (18.1 %). The place of occurrence for outdoor falls is on the street (5 %) and then on the highway (4 %).

The 2006–2007 annual cost of hospital treatment due to falls for older people aged 65 years and over was estimated around $600.3 million, an increase of $34.3 from $566.0 million in 2003–2004 [24]. A literature review conducted by Hauer et al. [25] found that there are three major ways to collect falls data: a retrospective system incorporating a postal questionnaire or an interview either by phone or face-to-face (n = 24, 27 %), a prospective system involving a postcard, a calendar or a diary (n = 38, 42 %), video camera monitoring or interpretation and analysis of health data record (n = 16, 18 %). The first two methods can be used as a gold standard for automated fall detection.

A research in fall detection system can be divided broadly into three different approaches: video monitoring system, body-worn and environmental sensors [26].

4.1 Video Monitoring System

One of the advantages of incorporating video monitoring system instead other approaches is the ability to capture different events that occur at the same time and place [27]. Furthermore, prices and installation costs for video monitoring system are cheaper and there is no requirement for wearing any form of sensors for 24 h a day [28].

4.1.1 Current Technologies in Video Surveillance

Leone et al. [29] made use of a 3D time-of flight vision technology to differentiate falls from normal daily activities. The proposed algorithm consists of three main features: the centroid height of the person, the duration of the critical stage, and the duration of the post fall stage. The threshold values are listed in Table 2. The algorithm validation was performed using 460 activity samples (200 normal activities and 260 simulated falls) from 13 healthy and middle-aged men (age: 30–40 years, height: 1.55–1.95 m). The simulated falls were performed on three different directions (backward, forward and lateral) and either incorporated or unincorporated recovery post fall. The combination of three thresholds was able to obtain an accuracy of 100 % when an occluded object was not obstructed from the view of camera. However, the system was only tested on healthy adults.

Shim et al. [30] attached web cameras to the patient's bed at The Severance Hospital of the Yonsei University Health System. First, rapid background subtraction from video sequences were performed to identify the entire movements of the subject and followed by counting the body width–height ratio to differentiate falls from daily activities. The algorithm assumes that someone might be fallen if a rectangular shape is detected for a long time outside the safety regions. The overall accuracy of the system was 93 %. No details of the participants were reported in this publication.

Table 2 The threshold values used in the proposed algorithm [29]

Threshold	TH1	TH2	TH3
Measure	Centroid height	Critical phase duration	Post fall phase duration
Unit	Meters	Milliseconds	Seconds
Value	0.4	900	4

4.1.2 Challenges

There are three common problems that must be considered when utilizing video monitoring system to distinguish falls from normal activities: a concern of privacy [31], high consumption of power and the complexity of data transmission networks [32].

A study conducted by Steele et al. has revealed that around 78 % out of older people would not accept the use of video monitoring system since it continuously records every event around the house and breaches their privacy [33]. A video monitoring system requires a large power input since it is operated on high frequencies and a continuous integration and clearance of charge must be maintained as light reaches a large pixel array [34]. A video monitoring system also suffers from long setup time, this condition occurs because the automation of auto-gain and white-balance and consequently is not suitable for low-duty-cycle networks. There is an issue with bandwidth capacity when raw images are directly transferred to central server. However, this issue can be solved by preprocessing the images in the local server and then only the meta-information will be sent to central server [34].

4.2 Body-worn Sensors

Since there is raised concern of privacy among older people in the use of video monitoring system for detecting falls [35], some current researchers have recommended the use of body-worn sensor to identify falls. The body-worn sensor approach is defined as an activity that involves holding or wearing embedded devices with the purpose of distinguishing falls from normal activities [28].

4.2.1 Current Technologies

Eight volunteers performed a series of simulated falls (forward, backward, and sideways falls) in order to test the system's performance that was made by Ojetola et al. [36]. The device consisted of a triaxial accelerometer, a triaxial gyroscope, a Bluetooth device and a microcontroller device. These devices were placed on the chest and the right thigh. Both the time stamp of each sensor node and the sensing data are transmitted to a remote station via Bluetooth for further analysis. The author

used a machine learning approach (decision tree classifiers) that combined with filtering (moving window size of 0.8 s) to differentiate fall events from normal activities. The associated precision and recall of the proposed system are 81 and 92 % respectively. Even though the preliminary results are quite good, but the placement of sensor nodes is not convenient enough to gain acceptance from the older people.

Tolkiehn et al. [37] placed a triaxial accelerometer, a barometric pressure sensor and a wireless transmitter into an enclosure. The device was worn around the waist of the subject. Eight males and four females were recruited to perform 13 fall events (falls from different furnitures and falls in different directions) and 12 normal activities. For safety purposes, the simulated falls were performed onto mattress. The fall events were recognized by comparing the changes of tilt angle and acceleration peak with predetermined thresholds. Although the associated accuracy was higher than 94 %, the resulted impacts may differ from the actual falls in older people.

Viet et al. [38] embedded an accelerometer and an orientation sensor into the Android GI phone and performed a laboratory trial that involved five young and healthy volunteers. Each subject performed three simulated falls (forward, backward, and sideways falls) and two normal activities. There were two thresholds used in the proposed algorithm: the changes in acceleration and orientation. The system indicated a fall may have occurred if the sudden acceleration and orientation changes of the mobile phone were more than the acceptable limits. The accuracy of the system was quite good (85 %) for identifying three types of falls, but the main issue of implementing this system was the fact that the user forget to carry their mobile phone.

4.2.2 Challenges

There are three major weakness of the system that incorporating body-worn sensor to distinguish falls from daily activities: uncomfortable feeling of such a sensor, experience labeling and stigma because of their age and physical disabilities, or simply suffering from memory problem due to dementia [39].

The acceptance of body-worn device among older people will depend mostly on the convenience and comfort when using them. A research by Stikic et al. [40] has stated that there is a tendency of lower acceptance since the device must be worn for 24 h a day for avoiding a gap of time in real-time monitoring. Algase revealed that most people suffering from dementia often forgot to wear such a sensor while doing their daily activities [41]. Aside from those drawbacks, the availability of push button on the body-worn sensor could be really helpful to contact the nearest emergency response center directly in a situation where a fall has occurred.

4.3 Environmental Sensors

There are two main advantages provided by environmental sensors: low or no privacy concern as compared to video-based approach [26] and more comfortable since there is no need to attach or wear a device for a whole day [42]. The definition of environmental sensor approach is the activity of placing multiple environmental sensors to fully monitor all activities that occur around the house [28].

4.3.1 Current Technologies

Litvak et al. [43] used a pattern recognition algorithm to identify a fall by fusing floor vibration and sound data. In the first stage, a mannequin called "Rescue Randy" was used to identify a forward fall at two different distances from the sensors (2 and 5 m) on two different floor surfaces (carpet and concrete floors). In the second stage, the same experiments were repeated by using different objects including a heavy bag, a plastic box, a book and a metal book. The associated sensitivity and specificity of the system were 97.5 % and 98.5 %, respectively. The limitation of the system was incapable in detecting a low impact fall-detection.

Liu et al. [44] mounted an array of passive infrared sensors on the wall at three different heights (50, 110 and 140 cm from the ground) and used two-layer hidden Markov model to distinguish falls. The system decided that a fall has occurred if the output waveform shows the characteristic of fall event. This system had an accuracy of 92.5 %. However, it should be noted that the trial only involved four healthy adults (one female and three males) performing forward falls onto a large mattress at distance between 1 and 3 m from the sensors.

Ariani et al. [45] used a simulated environment to understand the effectiveness of using motion and pressure mat sensors for recognizing three types of post-falls (a fall with recovery, a fall with or without loss of consciousness) from daily activities. Graph theory concepts are used to track the location of each individual and to monitor them separately for falls. This graph also used to determine whether someone has left or entered the residential. The proposed algorithm has achieved an accuracy of 93.33 % and 87.33 % for scenarios with one person and multiple persons. One of the major limitations in this system is that the environmental sensors are infinitely sensitive to any kind of movements. In the actual situation, the motion sensors are not activated or only generate signals for short period of time, when the object is moving away from the sensors [46]. Also if animals living in the residential, the system can not differentiate if the pressure applied to the mat is from the human body weight or an animal's weight, as a consequence, this approach tends to have a higher rate of false-positive results [47].

4.3.2 Challenges

Currently, the cost and system complexity still constitute the main issues for implementing the smart home technology, since a lot of sensors are required for monitoring comprehensively every activities in all the rooms of a home [45]. In the future, the technology will need smaller sensors with better energy efficiency; therefore the approach will be less expensive and widely deployed.

4.4 Discussion

The budget provided for the cost of device and maintenance must be reasonable when considering the expenses you would have to pay for the medical treatment of the fall-related injuries [33].

There is a big concern among older people about the sensor lifetime. They are worried that a fall may occur at the same time when the fall detection system is dysfunctional due to a dead battery. A research conducted by Noury et al. [48] has argued that the ideal battery life time should last for 1 year to avoid this situation from happening and to decrease the related maintenance costs.

A data collection from the South Australian Ambulance Service has revealed that the time gap between the first call to ambulance headquarters and the acceptable medical care would be around 5–15 min [49]. The Table 3 also shows that many of the patients are alone at the time of fall accidents. These facts show the importance to have a system that is able not only to automatically detect the occurrence of fall but also to track the subject's location. This function can be enhanced by embedding a feature to automatically detect falls in real time and sending an alert via mobile phone to the emergency contacts, including formal general practitioners and family member [50, 51].

5 Telehealthcare

A simple definition of telehealthcare can be stated as follows: "The remote delivery of personal healthcare services". First, physiological data such as blood pressure, pulse, weight and subjective patient data are collected from the patient. Then, data are electronically transmitted over the existing network infrastructure for remote examination by a medical practitioner. Finally, the medical practitioner provides suggestions based on the typical patient's current condition [52].

Telehealthcare can be categorized into three different systems: real-time (synchronous), store-and-forward (asynchronous), or hybrid systems [53]. In synchronous telehealthcare, an online real time communication is established between medical practitioner and patient and the acquired physiological data are collected and processed in real time, while in asynchronous telehealthcare the

Table 3 The occurrence rates of falls in association with the activation of emergency alarms for older people. Most of the users that were not holding or wearing embedded sensors were alone at the time when fall events occurred (64 out of 144 older people) and were not able to lift their bodies up from the floor by themselves [49]

		Alarm users (n = 124) n (%)	Matched sample of non-alarm users (by age, date and type of ambulance service, (n = 144) n (%)
Gender	Female	89 (72 %)	84 (58 %)[a]
Fall location	Home	121 (98 %)	100 (70 %)[a]
	Residential care facility	1 (1 %)	32 (22 %)
	Public place	0 (0 %)	12 (8 %)
	Not recorded	2 (1 %)	0 (0 %)
Alone at time of fall	Yes	97 (78 %)	64 (44 %)
	No	27 (22 %)	79 (55 %)
	Unknown	0 (0 %)	1 (1 %)
Ambulance called by	Self	115 (93 %)	21 (15 %)
	Family	5 (4 %)	66 (46 %)
	Health care staff/careers	2 (1.5 %)	41 (28 %)
	Other	2 (1.5 %)	16 (11 %)

[a] Significant difference between alarm and non-alarm users $p < 0.001$

acquired data are processed offline and both parties can view the results in their preferred time.

There are three main elements used in telehealthcare technology: portable medical devices, laptop, PC/laptop/tablet/smart phone, and a wireless network infrastructure [54]. The stability of the network is important when data is being transmitted over a network. Subsequently, the related data files are stored on a remote monitoring station for further analysis.

5.1 Telehealthcare for Chronic Diseases Management

Chronic diseases are the leading source of mortality and disabilities worldwide [55]. By 2020, it is projected that nearly 75 % of all deaths will be from chronic diseases. It is estimated that the death caused by cardiovascular diseases and hypertension to reach 16.7 and 7.1 million per year. The adult population with diabetes is projected to grow to 300 million people by 2025. Studies conducted by industry and academia have incorporated remote sensing to access and regularly monitor the health of older people. The availability of wearable/mobile health monitoring technology is really important since more than one-third of 78 million baby boomers and 34 million of their parents could suffer from chronic diseases including cardiovascular disease, stroke, etc.

The wearable health monitoring technology offered the following features [56]:

- To enable early detection of health deterioration.
- To send emergency notification to medical staff in the case of critical situations.
- To study the correlation between the health condition and lifestyle of an individual.
- To enable medical staff in monitoring the health condition of a person in a remote area through real-time transmission of physiological data.

5.1.1 Cardiovascular Diseases

Cardiovascular disease (CVD) is one of the world's top killer diseases [57]. Based on 2009 data from the American Heart Association, there are 80 million people with CVD in United States and around 2,400 Americans die every day because of CVD [57]. Cardiac arrhythmia, also defined as abnormal heart rhythms, is the most common type of heart diseases [57]. Healthy-looking people could suffer from arrhythmia and can go unnoticed [57]. The early detection of arrhythmia and their medical treatment are very crucial since arrhythmia can lead to a sudden cardiac arrest [57].

Some researchers have conducted studies using wireless monitoring system to continuously screen, diagnose, and manage patients who experience cardiac surgery and chronic heart failure. With this technology, cardiac implant patients can be remotely monitored no matter where the location. Cardiac data gathered from any patients can be seen by their cardiologist via mobile devices connected to the Wi-Fi/cellular wireless network [58].

The advantages that wireless monitoring system can offer to the cardiac patients are as follows [58]:

- Changes in condition for physically active patients can be assessed and monitored remotely which could lead to a decrease in the number of hospital readmissions.
- The patients can save time and cost for follow up care.
- A relatively lower risk for life threatening cardiac events can be achieved by continuously monitoring the condition of the patients.
- Patients in emergency situation are supported promptly and appropriately.
- Patients can have a better privacy and peace of mind while family members can provide extensive care to the patients. The heart's electrical activity, recorded as electrocardiogram (ECG) can be really helpful to detect the early symptom of poor blood flow, damage in heart muscle, abnormalities in heartbeats and other problems related to heart diseases.

Kim et al. [57] proposed a system that detects the symptoms of arrhythmia using ECG data collected with a portable ECG sensor. The sensor board includes an ECG sensor, a microprocessor unit (MPU) and a communication module. The MPU manages both the sensor and the communication module and also

responsible for processing raw data obtained from the sensor. There are three modules on MPU used for detecting the abnormalities in R–R interval data. The MPU also has two databases, the first one is used to store normal RR interval data, which is used as the gold standard for distinguishing an abnormal state in heart, while the second one is used to store the previous R–R interval data, which is used to recognize whether there is an increase or decrease in RR interval. The ECG sensor module measures the electrical signals emitted by the patient's heart every 5 ms. Then data are transmitted to MPU for further analysis. If the ECG sensor data contains abnormal heart rhythms and then the data are directly transferred to either a server or a handheld device via communication module. The accuracy of the system is more than 97 %.

Li et al. [59] proposed the use of a transductive transfer learning framework to recognize the abnormalities in ECG signals. The knowledge from labeled ECG signals were first used to classify unsupervised signals from target users. Then, the proposed method was applied to the MIT-BIH Arrhythmias Dataset and the performance results were compared with either anomaly detection or transductive learning baseline approaches. Overall, the results show that the proposed method has increased the system's effectiveness to identify the ECG abnormalities over the two methods for between 25 % and 60 % on G-Mean.

Gay et al. [60] developed a system intended for monitoring the health condition of patient with high risk of cardiac diseases. The system consists of wireless sensors, smart phones and data server. The physiological signals are measured by ECG, accelerometer, blood pressure monitor and weight scale. The collected data are transmitted to mobile phone for local processing and then the personal feedback is provided to the user. Both the specialists and the patients can assess the collected information through secure web portal. The specialists can remotely monitor and diagnose the patient's condition and suggest the necessary treatment. The initial trial involved two cardiac rehabilitation centers in Sydney (Australia) and one patient who recently experienced an acute cardiovascular event. At the end of the study, the patient was satisfied with the system and decided to continue the program.

5.1.2 Epilepsy

Approximately 60 million people in the world suffer from epilepsy [61]. An epileptic seizure is the result of an abnormal and excessive discharge of a set of cerebral neurons [62]. A tonic–clonic episode is the most common type of seizure experienced by people with epilepsy [62]. There are two stages involved in this tonic–clonic episode: a stage marked by tetanic muscular contractions (tonic) and then followed by a stage of jerking movements of the body and limbs (clonic) [62].

The anticonvulsant drug therapy is regarded as the most effective way to control approximately 67 % of seizure events with a further 7–8 % treated through neurosurgical interventions, which left 25 % of patients whose seizures could not be cured by any therapies [61]. The seizure management can be achieved through

medication and analysis of electroencephalography (EEG) and video data [63]. This approach provides a high accuracy of event detection and a low rate of false alarm. Unfortunately, this approach is not suitable for long-term support. A complex seizure affects memory of events and consciousness that may hinder the effectiveness of self-reporting of seizure incidence [64]. It must be noted that the treatment decision will be taken by physicians if the patient experienced frequent episodes of seizure attacks. Pharmacological treatment can be said successful if a patient has less frequent seizures after performing this treatment at certain time intervals [61, 65].

Dalton et al. [62] built a system for monitoring and identifying simple motor seizures. The system comprised of accelerometer-based kinematic sensors, a mobile phone and a base station. All of the trial's participants were patients at the Beth Israel Deaconess Medical Center. Each participant performed a series of activities including daily living activities (ADLs) and instrumental daily living activities (iADLs). Two parameters (EEG and video data) were recorded during monitoring session and used as a gold standard for detection of seizures. A template matching algorithm was developed to differentiate between seizure events and iADLs. The proposed algorithm was developed from a combination of a customized mass-spring template and the dynamic time warping in consideration that each person could experience different types of seizure events. From a total of 21 seizure events, the sensitivity of 91 % and the specificity of 84 % were achieved. Furthermore, the Mercury battery life was 10.5 h.

Shoeb et al. [66] evaluated the use of machine learning technique for developing classifier that intended for epileptic seizure onset detection. The evaluation is processed through the analysis process of the scalp recorded EEG. This work is quite challenging since the electrical activities of the brain are consisted of various classes with overlapping characteristics. The ways of obtaining high performance from the proposed algorithm are by applying the right framework for machine learning and by determining the critical points so that seizures can be differentiated from other brain activities. The accuracy of the proposed algorithm was 96 % of 173 seizures. This results was achieved when it is trained by data from patients who have two and more seizures and evaluated by 916 h of continuous EEG which are collected from 24 patients.

5.2 Hearing Impairment

Hearing loss is suffered by one in six adults in Australia [67]. One contributing factor to the increase of prevalence rate of hearing loss is the increase in age, increasing to less than 1 % for young children under 15 years to three out of four elderly people aged over 70 years [67]. The projected number of people with hearing loss will rise to one of every four Australians in the year 2050 [67].

One of the consequences from hearing loss is a reduction in ability for communicating clearly and effectively [67]. This consequence could be either mild (such as the reduction of hearing function) or severe (such as speech impairment)

[67]. It also affects the personal life of an individual through the lesser opportunity of gaining education, learning a new skill, finding a job and building a relationship [67]. Furthermore, the deterioration of health status is directly related to the loss of hearing [67].

Hearing can be defined as the softest and the loudest sounds that normal people can hear [68]. This hearing range is expressed in Hertz or number of cycles of a sound wave per second. The loudness level of the sound is between 0 and 140 dB, where 0 dB is the softest sound that a person with "normal" hearing function would be able to hear [69], while sound levels of 140 dB could lead to total loss of hearing.

Degree of hearing loss can be referred to the severity of loss (mild, moderate, severe or profound) [69]. The threshold ranges are different for people under and above 15 years [69]. The levels of hearing loss for a child under 15 years are: 0–30 dB (mild), 31–60 dB (moderate), 61–90 dB (severe) and over 91 dB (profound). While for a child aged over 15 years, the levels are: mild $\geq$ 25 and $<$45 dB, moderate $\geq$ 45 and $<$65 dB, and severe $\geq$ 65 dB.

There are two methods to evaluate the normality of cochlear function, subjective and objective tests [67]. An audiologist or a hearing aid audiometrist, or an ear, nose and throat specialist conducted a subjective test by performing an audiometric assessment. In adults, different types of sounds (include pure tones, speeches and noises from environment) are provided through the headphones. Those sounds are given in different level of loudness (decibels) and each individual reports the loudness of sound that they can hear by selecting the frequency range (usually the range is between 250 and 8,000 Hz). While in the young children, the specialist takes the measurement by observing the child's response to different sounds. In objective test, the auditory brainstem technique can be employed to measure the sign of hearing loss in newborn children. This approach can also be applied to people with the intention to determine the distributions of hearing loss configuration, the level of severity, and the site of lesion.

Dalton et al. [70] conducted a 5-year follow-up of 2,688 older people aged between 53 and 93 years to determine the association between hearing impairment and the quality of life. The study assessed communication difficulties by incorporating the Hearing Handicap for the Elderly Screening version (HHIE-S) and asking questions to understand the problem for each situation. The measurement of health related quality of life (HRQoL) is based on three main factors: activities of daily living (ADLs), instrumental ADLs (IADLs) and the Short Form 36 Health Survey (SF-36). The audiometric testing module is used to categorize the level of hearing loss. This module is based on the pure tone average (PTA) levels at 500, 1,000, 2,000 and 4,000 Hz. Research has revealed that people who suffers from moderate to severe hearing loss are most likely struggle to perform their activities (ADL OR = 1.54, 95 % CI = 1.06–2.24; IADL OR = 1.54, 95 % CI = 1.18–2.00). The results can be seen on Table 4.

The early assessment of hearing loss could prevent social isolation in young generation by delivering early intervention and rehabilitation. AL-Afsaa et al. [71] designed and developed a PC-based audiometer to detect hearing loss in newborns

Table 4 The association between the level of hearing impairment and participant's impairment ability to perform either ADL or IADL [70]

	No HL		Mild HL		Mod.-Severe HL		
	n	%	n	%	n	%	P value [a]
Impaired ADL							
All	92	7.3	108	14.9	174	27.9	<0.001
52–59 years	20	4.5	7	7.6	3	7.3	0.22
60–69 years	32	6.1	19	8.1	11	9.8	0.12
70–79 years	33	12.6	47	16.3	54	24.3	0
80–97 years	7	18.4	35	31.5	106	42.6	0
Impaired ADL							
All	436	34.6	338	46.7	359	59.1	<0.001
52–59 years	123	28.0	25	27.2	12	29.3	0.95
60–69 years	177	33.8	84	35.6	52	46.9	0.02
70–79 years	118	45.2	161	55.9	122	55.5	0.02
80–97 years	18	47.4	68	63.0	173	73.3	0

Note ADL: activity of daily living; IADL: instrumental ADL; HL: hearing loss
[a] This is by using of a Mantel–Haenszel Chi square test for trend

and young children. The system employed the auditory brainstem response (ABR) approach. Instead of incorporating a traditional click stimuli, a new approach called chirp stimulus was used to generate ABR waveforms. The chirp stimulus can be defined as a transient sound, where the lower frequencies may reach the participant's ear first since the higher frequencies are delayed, producing a very short tone glides in upward direction (5 ms) [72]. Moreover, the researcher designed a chirp stimulus with the idea of synchronizing the movement of hair cells in the cochlea. The final aim of this research was to develop a system to detect the sign of hearing impairment in early infancy and to perform hearing test that incorporating stimulus at low frequency. Currently, the development of intended chirp stimulus has been successfully completed and the ABR signals can be amplified by the implemented module to a recordable level. Furthermore, the implementation of the PC-prototype interface has been extended by using a Ni-DAQ data acquisition board.

A software called "AudioGene" has been developed by Taylor et al. [73] to identify the genetic causes of a certain form of inherited hearing loss–Autosomal Dominant Nonsyndromic Hearing Loss (ADNSHL). The associated software works based on machine learning techniques employing phenotypic data extracted from audiograms. When incorporating Multi-Instance Support Vector Machine (MI-SVM), the AutoGene was showing an estimated accuracy of around 68 % for predicting genes associated with ADNSHL (DFNA9, DFNA2, DFNA8/12). However, the estimated accuracy is quite low (approximately 44 %) when employing a Major classifier.

5.2.1 Musculoskeletal Conditions

One-third of Australian people are suffering from musculoskeletal conditions [74] and contributed to nearly quarter of all chronic diseases after experiencing an injury [75]. The patient needs to directly contact the physical therapist to get effective assessment and treatment of musculoskeletal conditions [76]. However, this rehabilitation service is quite limited in rural and remote areas because of the relatively small number of physical therapists [77].

Based on the 2004 data, the ratio of physical therapists per 10,000 population in Australia was nearly twice for capital cities (6.14) compared to 3.65 in remote areas and 3.58 in outer regional areas [78]. Based on the 2008 data, the ratio was worse (3.07) in New South Wales rural areas and could be dropped even further because of the recruiting and retaining issues and the fact that the professional staffs are getting older [79].

Russel et al. [77] developed the eHAB telerehabilitation system that allows the diagnosis and monitoring of health condition of patient with lower limb injury. The system consists of a measurement device and a video conference facility. The measurement device, which is powered by battery, enables the quantification of physical performance including balance, joint range of motion, muscle strength and gait assessment. The video conference facility allows health professionals to conduct re mote consultation and closely monitor progress of each patient in their own homes. The gathered data are transferred in real time using wireless communication. Nineteen participants were recruited for a two months trial. In this trial, a one and half hour session was conducted over a distance between the participant and the physical therapist. Each participant performed a series of activities, including a patient interview, a face-to-face physical examination, and a physical self-examination. The results of physical examination revealed that there was a significant agreement ($0.61 < k < 0.80$) in validity study and higher agreement ($0.81 < k < 1.00$) in the intra-rater and inter-rater reliability studies.

5.2.2 Respiratory Diseases

Three common types of respiratory diseases worldwide include: obstructive sleep apnea (OSA), chronic obstructive pulmonary disease (COPD) and asthma [80]. OSA is a kind of disorder in which breathing frequently stops and starts during sleep [81]. There are more than 18 million Americans diagnosed from OSA and it is projected that 10 million people with OSA remain undiagnosed and untreated [82, 83]. COPD is a serious lung disease, it will become the world's fourth leading cause of death, increased from the sixth place ranking in 1990. These situations are affected with the increase of smokers and the changes in demographic in many countries [84]. Asthma is a disorder characterized by blockage or narrowing of lung airways which lead to wheezing, shortness of breath, chest tightness, and coughing. Around 300 million people worldwide affected by asthma in 2010. Globally, the report revealed that about 25,000 deaths were linked to asthma in 2009 [85].

Cao et al. [80] proposed a respiratory monitoring system, which includes a micro thermal flow sensor for monitoring respiratory airflow, a triaxial micro accelerometer for monitoring body posture, and a micro-photoelectric sensor for monitoring blood oxygen saturation. The flow sensor is capable of detecting infants' respiration since it is sensitive enough for sensing weak airflow. This capability is achieved by using three defined thresholds: a threshold of 0.6 L/min (preterm infants and newborns) or 1.8 L/min (infants beyond the neonatal period) and a range of 0–10 L/min. Collected data are transmitted through a PC or a mobile phone via Bluetooth wireless to a remote server. Then data are analyzed to determine whether severe respiratory symptoms are present.

There are two main thresholds involved when decided whether or not sleep apnea exists: a respiratory parameter AHI (apnea-hypopnea index) and oxygen saturation. AHI can be defined as the total episodes of slow breathing (apnea-hypopnea) per hour of sleep. If AHI is more or equal to five in airflow or tidal volume and oxygen saturation is less than 90 %, then it is confirmed that the sleep apnea does exist.

The proposed system uses three different thresholds: forced expiratory volume after 1 s (FEV1), forced vital capacity (FVC), and forced expiratory flow at 25–75 % of maximal lung volume (FEF25–75) for detecting the occurrence of COPD and asthma. The patient is diagnosed with COPD or asthma if the ratio of FEV1/FVC is less than 70 %.

Corral et al. [86] proposed a system to be used for diagnosing and treating patients with obstructive sleep apnea syndrome (OSAS). The system comprises a polygraph, a webcam and a PC. The patients' data including oral/nasal flow, a pressure cannula, chest and abdomen movements, heart rate, body position, snoring, and oxygen saturation (SO2) were collected and then transmitted to remote server for further analysis. Forty patients (mean age: 53 $\pm$ 10.3 years, a body mass index (BMI): 31 $\pm$ 6.2 kg/m^2 and an Epworth score: 12 $\pm$ 5.3) participated in this study and a half of them were treated by conventional consultation while the other half were examined via teleconsultation. The researchers also measured the end user satisfaction level and the system cost. At the end of trial, OSAS were found in 35 people, with an Apnea-Hypopnea Index of plus or minus ten and then 16 of them were started with continuous positive airway pressure (CPAP) treatment. The correlation coefficients between data transmitted in real time and data stored in the polygraph amounted to 0.94 for total apneas/hyponeas, 0.97 for apnea-hyponea index, 0.98 for time of oxygen saturation and 0.97 for mean of oxygen saturation. The compliance rates reported for CPAP treatment with a conventional clinic and CPAP treatment with teleconsultation were 85 % and 75 %, respectively. The cost for implementing a telematics polygraph was estimated about 227 €.

5.2.3 Stress

Stress can cause physical (e.g., headaches, fatigue, sleeping difficulty, and heart diseases) and mental illnesses (which may result in decreased social activities and

social isolation) [87]. The stress can be managed if reliable monitoring and assessment of stress can be conducted in real environment, for example by using mobile applications. So far the stream of research has been focused on proposing a wearable device that can effectively monitor and assess stress in natural environment; this research can be quite challenging since it had not been scientifically proven yet [88].

Ertin et al. [89] proposed the use of two different models for identifying acute stress in individual based on the analysis of physiological signals. The physiological classifier was chosen for predicting the correlation between changes in physiology and stress. The output of physiological classifier was used as an input to the perceived stress model. This model was used for predicting the pattern resulted from the gradual and steady accumulation of stress in the human's mind.

The signals were captured by six different sensors: (a) the heart's electrical signals were measured using an electrocardiograph (ECG) by placing two electrodes at specific areas on the body; (b) the skin conductance was measured by ECG electrodes; (c) a skin temperature thermistor placed on the skin mid thorax; d) an ambient temperature sensor; (e) a triaxial accelerometer; (f) relative lung volume at the rib cage was measured by a respiratory inductive plethysmograph (RIP) band. The collected data from each sensor were transferred to a mobile phone through Bluetooth communication.

The trial recruited 21 subjects from University of Minnesota, Duluth. Each participant performed three main sessions (public speaking, mental arithmetic, and cold pressor) in a laboratory setting. The reasons behind the choice of activities were because those activities represented the social, cognitive, and physical stressors that occurred in daily life and the most known factor for inducing stress among people.

There were 14 questionnaires to be filled during the lab session with the intention for capturing not only subjective stress but also non-stress states. The synchronization between the questionnaire response and the physiological data was critical for optimal correlation. There were five questions stated in the questionnaire list: cheerful? happy? angry/frustrated? nervous/stressed? and sad? A four-point scale was used for each question: 0 (NO), 1 (no), 2 (yes), and 3 (YES). The session will be rescheduled if the participant consumes any caffeine (drink caffeine-free sodas, coffee and tea), smokes, and performs strenuous exercise at least four hours before the session is started, or drinks alcohol the day before the session and consumes painkiller pills at least three days before the session is conducted.

Findings revealed that the 90 % accuracy was reached for the physiological classifier while the median correlation with self-reported rating was 0.72 for the perceived stress model.

5.2.4 Stroke

According to 2009–2010 data from the U.S. Department of Health and Human Services, stroke has become the fourth biggest killers in United States [90, 91].

The projected number of Americans with stroke was around 7 million in 2008 [92]. The U.S. national cost of stroke in 2008 was around $18.8 billion, including nearly 50 % of total medical cost for hospital services [92]. There are health campaigns to educate the public on decreasing the risk of stroke by encouraging them to keep physically active, stop smoking and take medication for lowering blood pressure and cholesterol [93, 94].

Stroke can cause motor system dysfunction [95]. This dysfunctional can be cured by performing rehabilitation exercises regularly to improve patients limb function, posture and gait [95]. Ma et al. [95] proposed a system that enables continuous monitoring and evaluation of patients with post stroke. The architecture of the system including:

- A kinematics node, which contains a triaxial accelerometer, one gyroscope and a wireless transducer, placed in three different body parts (thigh, knee and ankle).
- Each insole contains an eight-channel plantar pressure measurement device and a wireless transducer used to convert the output data.
- A wireless sensor network gateway.
- A computer software module created to provide the interaction between the patients and the staffs in rehabilitation department and collected data can be sent to remote server through the Internet.
- To maximize the effectiveness of man–machine interactions, the system was proposed to also include multimedia devices.
- A remote server was planned to be placed in the nearest hospital's rehabilitation department.

The researchers were still working on the ongoing implementation of the proposed system (Table 5).

5.2.5 Sleep Problem

Sleep is the most important factor in either physical or mental health and about one-third of people's lifetime is spent for sleeping [96]. Sleep problems are the most disturbing among older people. In a research involving more than 9,000 older people aged 65 year and over, more than half of them were reported insomnia, difficult to wake up or wake up too early, or a need to take a nap and not getting enough rest.

In fact, older people with sleep disorders have a cumulative effect on their physical and mental health and these disorders can cause decrease health function, memory problems, greater risk of falling and higher risk of death. Lifestyle-related sleep disorders in older people can also be associated with other health problems, such as congestive heart failure, cancer, nocturia, breathlessness caused by chronic obstructive pulmonary disease, the deficiency of neurons derived from cerebrovascular accident and Parkinson's disease [97].

Table 5 The types of motor training, the appliances and the sensors [95]

Motor training types	Appliances	Sensors	Parameters for evaluation
Upper extremities reaching and grasping	Upper extremities training robot	EMG electrodes, accelerometers	EMG electrodes, accelerometers
Spasticity relief	Functional electricity stimulation (FES) devices	EMG electrodes	EMG electrodes
Walking training	Canes, exoskeleton frame	Accelerometers, gyroscopes	Accelerometers, gyroscopes
Deformation relief	Hand/foot orthoses	Pressure sensors	Pressure, posture

Ni et al. [98] proposed a sleep pattern monitoring system that incorporates pressure sensor matrix. A total of ten students (eight men, two women; ages 21–26 years) were recruited for this trial. In this trial, the participants performed three kinds of sleep postures: left-lateral sleep, right-lateral sleep and supine sleep. When pressures were applied to the Flexi-force sensor, the analog signals were sampled and converted to digital data. At the end of trial, 150 datasets were collected. Then to obtain the information about elder's sleep behavior in a bed, the data were analyzed using two classification methods, such as naive bayes and random forest. Next, the experimental results were validated with two validation process, 10-fold cross validation and leave-one-out cross validation. The output from the analysis results can be viewed by both the user and their caregivers. In addition, the system can give a feedback about the best sleeping position according to their disease history and symptoms.

The highest average accuracy was 87.33 % when using a combination of random forest and leave-one-out cross validation. Although the results of the trial was quite promising, further trials which include older people will be needed to test the proposed system.

5.2.6 Tuberculosis (TB)

According to 2011 data published by World Health Organization (WHO) [99], around 8.7 million new cases of TB were reported, where 13 % of the patients were co-infected by the HIV virus. Furthermore, around 1.43 million people died from TB, of which 30 % are found among people with HIV. TB affects women more than men, where 0.3 million HIV-negative women and 0.2 million HIV-positive women died because of this disease.

In the first stage, Koesoema et al. [100] proposed a preliminary model of system intended for a TB treatment and control management, especially in countries in development process. The research's goal is to provide a framework for integrated information that can be useful for assisting the treatment of tuberculosis based on Directly Observed Treatment Short-course (DOTS). The preliminary design

system with software modules are expected for implementing the following functions: "(i) patient treatment surveillance data recording and reporting, (ii) general epidemiology and statistical data recording and reporting, (iii) case finding, diagnostic tool and laboratory examinations, (iv) medication-and-supplies inventory and distribution, (v) personnel training as well as (vi) community education and environment conditioning."

Various telecommunication networks available at key locations (such as internet, Global System for Mobile communications (GSM) and phone line) were used for the transmission of healthcare data, the diagnosis of patient's health status from the medical staff, the management in the distribution of medication and teleeducation. The researchers have a plan for conducting the pilot trial in local health office in Bandung and community health centers. The DOTS program has also been implemented in both centers. It is hoped that in the future, this design will be integrated with other healthcare projects.

In the second stage, the preliminary development of telehealthcare for managing tuberculosis in either a community health center (CHC) or a hospital has been established by Rachmat et al. [101, 102]. The diagnosis and the therapy for tuberculosis patients in Indonesia are based on the DOTS strategy. However, there is a high prevalence of tuberculosis in Indonesia due to incomplete therapy and follow-up. There are a couple of reasons why patient stops their therapy including feeling much better, forget to take their medication on time, not regularly visit the doctor for further follow up and possible suffer from the side effects of medications.

There are five main components of TB management e-health system: personal computer (PC), simple digital microscope, patient database software, short message services (SMS) gateway software, and telecommunication module.

First, the sputum samples are captured using a digital camera. Next, the examination of sputum for detecting the availability of acid-fast bacilli (AFB) is performed by using a digital microscope connected to a computer. Then, the TB diagnosis quality improvement could be reached by continuously evaluating the data that have been transferred to central database. Within six months, 16 SMS reminders could be sent on a specific date to inform the patient about the medication schedule or the doctor appointment. Furthermore, the remote consultation can be performed through SMS facility on mobile phone.

The system can be extended by adding a facility that can be used for a remote consultation between the patient and a pulmonologist or other medical specialist in particular hospital. In general, there are two goals that researchers hope to achieve: the increasing number of TB patients who completed their therapy treatment and the improvement of quality in health care services.

Histogram based statistical features and evolutionary based extreme learning machines were chosen by Priya et al. [103] to automate the decision support process for differentiating TB positive and negative on digital images. First, the images which consisted of sputum smear both positive and negative outcomes were recorded by using standard image acquisition protocol. Next, those recorded images were processed by histogram based feature extraction technique. Continue with the process of finding the most significant features using student 't' test.

The selected features were then used as input to the differential evolutionary extreme learning machine classifier. The final result has proved that the histogram based significant features are able to distinguish TB positive from TB negative with the accuracy of 96.6 % and 92.5 % for training and validation, respectively.

5.3 Telehealthcare for Health and Physiological Monitoring

Another innovation in wireless technology has enabled the development of sensor system that is intended to monitor the person's health condition and their physical activities [104, 105]. This wireless technology offers comfort and convenience for the user since no cable is required to transmit data [104]. The wireless body sensor networks/systems have grown into one of promising technologies for monitoring the health status of an individual [106–110].

Huo et al. [13] have developed a smart home for older people using the combination of wireless body and home sensor networks. In addition, the system also completed with other functionalities including outdoor monitoring, and emergency decision and alarms. The body sensor network may be composed of temperature and pulse sensors, accelerometer and even ECG, eletromyography (EMG), EEG etc. It depends on the health condition of each older person. The home sensor network may consist of many different types of sensors installed in fixed locations. They include temperature, relative humidity, air pressure, light sensors and microphone in different locations of the apartment, force-sensitive resistors in chairs, sofas and beds. When the person stays inside the house, the data collected from body sensor network are transmitted to the central server via Home Sensor Network (HSN) gateway. When the person leaves the house, the data collected from body sensor network are sent to central server via public communication network, such as GSM, 3G-WCDMA, Wi-Fi, x-DSL, Ethernet, and future information networks, e.g., NGN/IMS. The telephone call or SMS or email will be generated automatically and sent to the caregivers and the family member when the system detects emergency situations. In this proposed system, the implementation of interconnection network is significantly important to establish multimodal interactions with communication devices. There are three key issues that need to be addressed: hand-off the Body Sensor Network (BSN) from HSN to public networks and from public networks to the HSN; multi-network communication with the same standard structure; the network can be expanded with next-generation network (NGN) capabilities.

Lv et al. [111] proposed a system consisting of the following four main elements: various body sensors and medical devices, smart phone, server, and emergency center. As the first elements, the researchers used off-the-shelf devices, including an integrated Bluetooth ECG/Accelerometer sensor from Alive Technologies and a Bluetooth enabled blood pressure. For the second element, a smart phone is used to analyze the physiological data from the sensors and then to deliver the results to the central server. The smart phone is also capable of sending emergency message to hospital and the family members when an emergency

situation is detected. The third element is a central server that serves as a database and provides real time medical consultation directly to both the patients and their family members. The emergency center is the fourth element, to provide medical support by calling an ambulance to the emergency scene, upon receiving an emergency notification through GSM protocol.

Wu et al. [112] has developed a system that combines a wearable personal healthcare and emergency aid system. The users' physiological signals are continuously collected with small wearable sensors and then the collected data are transferred to a mobile phone via Bluetooth. The mobile phone has three important roles: data storage, processing and analysis data, sending a periodic report of health status of the user to the healthcare center through GSM module and generating alert to call medical staff in case of emergency. However, the proposed system only offers one-way communication that is not sufficient enough for providing real-time and dynamic monitoring.

Prognosis is a monitoring health system that enables the fusion of physiological signal [113]. The wearable system has been integrated with decision support system (DSS) and employed finite state machine operations for attempting to correctly detect abnormal events. Moreover, it is also able to track the health status of the user and to generate emergency signals.

Chang et al. [114] developed a smart chair called "e-caring" to measure the blood pressure, body temperature, heart rate, height, weight and body fat percentage, and then the results and the diagnoses are real time reported to the user. This report is generated by the e-caring after the system performed the point and trend exception detecting procedure. The green light is displayed if the user condition is healthy; the yellow light is shown if user's condition deteriorates, while the red light for the user who has shown symptoms of an illness. When the report light is yellow, the user is recommended to fill out a predefined questionnaire which will be used as self-assessment and reference. However, if the report light is red then the user is advised to see their general practitioner (GP) for further evaluation. The report can be sent automatically to their GP if the corresponding email address is available in the system. Due to privacy concerns, both the users and their GPs can only access the health information through secure web portal in their own place.

Regular physical activities could contribute to the increasing level of fitness and the capacity of exercise, decreases the risk of having health problems such as obesity, diabetes, blood pressure, and lipid abnormalities, enhances the cardiovascular function and increases life expectancy [115]. This system works by measuring the energy expenditure of a person in a real time [116].

Mo et al. [116] designed a system that can be used to assess the physical activity of an individual. The system incorporated two different sensors: two triaxial accelerometers and one displacement sensor. The accelerometers at the hip and wrist sites are used to measure the whole-body and arm movements and the collected data are used for the characterization of degree of physical activity (PA). The displacement sensor, placed around the abdomen area, is used to measure breathing rate and volume. There are two network topologies used for this study: (a) the hip unit

received signals from the abdominal (AB) and wrist unit through ZigBee module and then a 2-GB digital micro SD card embedded in the hip unit is used to store the received signals. (b) The nearest master node (MN), such as a smartphone or computer, received signals from the three sensing units. This study recruited three males and five females and the participants' characteristics are (mean ± standard deviation): age = 25.9 ± 7.0 years, mass = 70.0 ± 12.9 kg, height = 166.8 ± 12.8 cm, and body mass index = 25.2 ± 3.7 kg/m^2. Each participant is requested to perform a series of daily activities, for example, computer work, driving to work, walking stairs, life style activities, etc. The session itself will take around two h. The participant's activities are observed and monitored via palm device by a trained researcher. The trained researcher tried to have minimal contact with the participant with the intention of not disturbing the participant's activities. The support vector machine (SVM) classifier achieved an accuracy of 86 % for predicting the level of activity intensity.

Aridarma et al. [117] designed and developed a system capable of measuring six types of physiological signals, including ECG, temperature, body-mass index (BMI), heart and lung sound, and blood pressure. The proposed system consists of a mobile device and "off-the-shelf" sensors. Each measured signal is time-stamped when it is stored to the database. A knowledge base is designed to analyze the physiological signals and to distinguish any signs that the patient's health is deteriorating. The stored data can be used as preliminary results before the doctor proceeds to do the medical check-up. The benefits of the proposed system are accelerating the prevention of health problem and promoting the sustainability of lifestyle preventive interventions.

Sutjiredjeki et al. [118] developed a mobile telemedicine system with multi communication links that can be used in either normal situation or emergency event such as disaster. The developed system is consisted of two main units, a mobile telemedicine unit (as shown in Fig. 1) and a base unit.

A mobile telemedicine unit is packed in a case of 0.46 × 0.16 × 0.32 m and consists blood pressure, fetal heart rate, and 12 lead ECG monitors, thermometer, and blood oxygen saturation (SpO_2) units, as well as a telemedicine arbiter. The kit is also equipped with a core 2 duo processor, a 1-GB memory, a 3-megapixel digital camera, a 12 volts power supply and communication devices. The collected data can be transferred through various communication networks including POTS, GSM, CDMA, VHF radio, internet, and satellite [119]. Since this kit is also intended for emergency event, the GEOSAT FR-190G terminal can be embedded into the kit. This terminal is connected to ASIA Cellular Satellite (ACeS) network that provides services for customers within Asia Pacific region.

The base unit is consisted of a PC server with the following specifications: processor Core 2 Duo, memory DDR II 1,024 Mb, 80 GB hard disc, 2.2 GHz clock frequency. A database on the server can be used for patient information storage. The server also uses the same communication devices compatible with the mobile telemedicine unit.

At first, the system is only intended for the improvement of health services in the remote area [120]. But then, the design of the system is enhanced for providing

Fig. 1 A mobile telemedicine unit

healthcare services during a disaster. The system can be used for recording and reporting the patient's vital signals, teleconsultation, telediagnosis, teleeducation and the monitoring of health status of the patient.

5.4 Discussion

An extensive study was conducted to identify the issues relating to the use of sensors for home healthcare and emergency [121]. This implementation has raised a number of potential challenges including interoperability and interference, real-time data acquisition and processing, reliability and robustness, energy conservation, data management, data privacy and security, and comfort and unobtrusive operation. Furthermore, the author also pointed out the design issues when implementing this system including sensor/hardware development, mobility of wireless sensor, cost and size of the wireless node, network infrastructure, network coverage, network size, power management, life-time of the sensor networks, and quality of service requirements in such networks.

Huang et al. [122] conducted a study of acceptance of telehealthcare among Taiwan residents (males (n = 173), females (n = 196)). The forecasting model in telehealthcare was analyzed using the structural equation modeling (SEM) technique. The findings revealed that the model is capable of describing the behavioral interests for using telehealthcare. The top five factors affecting the consumer

interest were: attitude, the perceived usefulness, the perceived ease of use, subjective norms, and personal innovativeness. The results revealed from this study can be useful for reference purposes.

5.4.1 Quality of Service (QoS)

The implementation of QoS in WSN is a big task that quite challenging [123]. Sensor devices are scarce resources because they are low in battery power, have enough memory storage capacity, the capability to process data and offers high-speed wireless connection. There are a lot of parameters to consider during the design of WSN with QoS. A number of user requirements regarding the network are listed below:

- Data accuracy: the precision and the accuracy of data that have been collected.
- Quality: quality of video files, smoothness of video stream, detail of picture, etc.
- Timeliness: low delay data transmission.
- Reliability: the ability to correctly identify events and then store the collected data to database.
- Context-awareness: the precision and the accuracy of data that have been collected.
- Interoperability: the ability to connect with other systems (GSM network, radio-frequency identification (RFID) tags, Internet, etc).
- The lifetime of network: the time remaining before node is considered dead or the battery life.

5.4.2 Acceptability

Milhailidis et al. [124] conducted a study to measure the acceptability of monitoring technologies among baby boomers and older adults. The researchers recruited 15 baby boomers (born between 1946 and 1965; ten women and five men) and 15 older adults (born before 1946; eight women and seven men) which meet the inclusion criteria, includes the participant must live within the greater Toronto area, fluent in English, and no history of cognitive impairment.

The aims of the study were as follows:

- Identify the preferred types of sensing and monitoring technologies.
- Identify the ideal location for the placement of the sensors.
- Identify the principal factors that can affect the motivation and willingness to use those technologies in daily life.
- Compare the willingness to use those technologies between the baby boomers and the older adults (Table 6).

Findings from the research revealed that both generations have agreed to use the technologies as long as they can remain independent and safe in their own home.

Table 6 p values for Chi square tests calculated from the acceptability of installing the sensing and monitoring technologies in different locations around the house [124]

	Location						
	Bedroom	Bathroom	Kitchen	Living room	Hallway	Outdoor	NA
Type of monitoring technologies							1.000
PERS							
FD	0.682	1.000	0.682	0.682	0.682	0.139	
EC	0.500	0.100	0.224	1.000	0.042*	0.169	
LM HPMA	1.000	0.715	1.000	1.000	0.715	0.462	0.224
PMT	1.000	0.427	1.000	0.466	0.466	0.466	
HPMB INT	1.000	1.000	0.715	0.462	0.264	0.450	1.000
Type of sensors	0.042*	0.100	0.100	0.006*	0.017	<0.001*	1.000
SW							
MS	0.330	0.169	0.330	0.330	0.042*	0.080	
VC	0.450	0.700	0.715	1.000	1.000	0.710	
CV OPS	0.700	1.000	1.000	1.000	1.000	0.450	1.000

Note PERS = personal emergency response system; FD = fall detection; EC = environmental controls; LM = lifestyle monitoring; HPMA = health and physiological monitoring type A; PMT = ADL Promptin; HPMB = health and physiological monitoring type B; INT = interactive video teleconferencing; SW = switches; MS = motion sensors; VC = video cameras; CV = computer vision; OPS = on-person sensors; *$p < 0.05$; Findings from the research revealed that both generations have agreed to use those technologies as long as they can remain independent and safe in their own home

5.4.3 Data Accuracy

Some studies focused to monitor the patient's condition by using a wearable sensor to measure one physiological parameter and to generate alerts if there is a symptom of deterioration [125]. However, this approach may be less accurate than when used of more than one physiological parameter and combined with the patient's context. The diagnostic accuracy for wearable sensor system can be improved by adding contextual information, including the activity of the patient, environment, and social interaction [126]. This step is chosen to find the correlation between the physiological and contextual data. For instance: the health condition of the patient with heart diseases can be categorized as normal even though the ECG sensor has detected an increase in heart rate if the activity sensors (e.g., accelerometers) have noticed that the patient's movement is faster than its average motion due to extensive exercise training [127]. In this case, the contextual information of the patient consist of the patient's static posture, dynamic activity, orientation, gestures, biometric, and emotional states. While for the environmental context comprises the patient's location, time, humidity, barometric pressure, illumination, and noise [128]. Followed for the social interaction includes the face to face communication with friends or family around the environment

[129]. Moreover, further information about the state of the patient can be gained by incorporating different types of sensors [130].

5.4.4 Data Privacy and Accuracy

There are serious concerns for secure transmission and storage of health information, when considering the importance of privacy of health information. Telehealthcare requires a well-established protocol and a proper regulation for encouraging people to use a health monitoring system in their daily life [131]. Kotz et al. [131] tried to understand the weakness of the existing privacy frameworks, found every possible solutions with the intention to develop a robust framework for mobile healthcare system and followed by proposing the implementation of their system based on the outcomes of the research. The researchers also brought the privacy issues and questions that need to be reviewed by all related parties including technical personnel, governmental bodies and regulatory organizations.

Simpson et al. [132] discussed the issues of home monitoring system including which parties can access the information; what information should be included without breaching the privacy and security; how the information is presented since different user will choose their communication styles such as telephone, computer, etc. Cryptography [133] and context-aware access control rules [134] are two such methods currently proposed as solutions for protecting the security of data.

5.4.5 Reliability

Reliability can become a critical concern when using telehealthcare in residential environment. The wireless network deployed in a hostile environment might suffer from many reliability issues [135]. Findings from a study conducted by Fong and Pecht [135] in rural China have revealed that the four factors affecting the network reliability were physical obstacles, atmospheric absorption, inadequate fade margin and system failure.

6 Medical and Medication Intake Management

A report has revealed that nearly 55 % of American senior citizens have failed to regularly take their medicine and 26 % of errors make the condition worse [136]. Therefore, the medical staffs including doctors, nurses or other caregivers must provide constant care to the senior citizens [137]. Consequently, more and more research are focused on monitoring medication intake with the intention to remotely maintain the health conditions of senior citizens [137].

6.1 Medical Ventilator

In critical care unit of a hospital, a patient with respiratory difficulties will need medical ventilator to help with breathing [138, 139]. This device mechanically supplies oxygen (O_2) into the lungs and moves carbon dioxide out of lungs. There are four main components of medical ventilator: air and oxygen supply tanks, a compressible air reservoir, a variety of valves and tubes, and a disposable or reusable patient circuit.

The compression of air reservoir is performed several times per 60 s to transfer a mixture of air, oxygen or other gases to the patient. The oxygen concentration in inspired air is around 21.

Cheng et al. [139] proposed a programmable medical ventilator which can be beneficial for the patient itself and medical staffs. It is more comfortable for the patient since the air flow and pressure can be dynamically adjusted based on the patient's need. It is more time-effective since the medical staffs can remotely monitor the patient's respiratory status and sets the operation of medical ventilator. There are four parameters that are monitored by the ventilator:

- Patient-related parameters including the flow, pressure, and volume of the air. These parameters are measured with sensors attached to a patient.
- Ventilator functions such as power failure, air leakage, pneumatic unit, and mechanical problems.
- Backup power supply.
- Oxygen tanks.

In the initial stage, the minimum level for respiratory rate and gas volume is manually set by the medical staff. Then, the medical ventilator will automatically maintain the normal beats per minute (BPM) rate as there are changes in change in the pressure of the lung. A healthy person has a normal BPM of 12.

6.2 PCA Infusion Pump

A cyber-physical system for modern hospital integrates various medical devices over the network. One of the examples involves the use of patient-controlled analgesia (PCA) infusion pump to manage the pain after surgery [140]. Each patient may react differently to a medication and the administering of medication must be based solely on the patient's condition. The dosage could be decreased or increased by pressing the button on the pump. However, this self-administer could lead to overdose which can cause fatal consequences such as respiratory failure.

As a solution, the dose of opioids must be injected and maintained at correct level by a programmable PCA system so that overdoses can be avoided, regardless the button is being pressed frequently by the patient. But still there is another weakness in this safety mechanism. Errors in programmable pump or a man-made

mistake when changing the dose of opioids could lead to overdoses because the drug concentration loaded in the pump is incorrect. This means the researcher needs to find another solution to address all clinical practice issues since the existing safeguards, for example drug libraries and programmable limits, can only be used for certain cases [141].

Pajic et al. [140] proposed a system that enables to detect the early signs of respiratory failure and followed by stopping the pump and notifying the medical staff that the adverse event has occurred. The system comprises of pulse oximeter which is used for evaluating the value of SpO_2 and heart rate of the patient. The critical values for oxygen saturation and heart rate are 70 % and 11.5 beats per minute, respectively. The final results have shown that the system was able to detect 22 cases of respiratory failures.

6.3 Personalized Medicine

Mathews et al. [142] developed a smart device called "MD.2" to remind older people to take their medication regularly. The size of MD.2 is about the size of coffee maker and the medications can be stored in it for up to 4 weeks. The reminder alert can be set to three different ways: via beeping, flashing light, or a voice command and also completed with special messages. A staff in remote support center can assess the medication profile, including the dosage and the schedule, and can be contacted by telephone, fax, or internet. The system will notify the caregivers or health professional if 90 min has elapsed and the older people are still not aware that they completely forgot to take their medication in time. The system has the ability to record the history of pills that have been dispensed to older people and to track adherence for a period of 5 weeks.

In Range has introduced the system called "EMMA" to remotely monitor their condition and adhere to their medications [143]. The system has been approved by the Food and Drug Administration (FDA) and is classified as FDA class II medical. The medical staff can make changes in medication dosage and frequency from remote center. This step is effective to decrease the number of visits by caregivers or medical staffs. Blister packaging is used to store the medication and the audio or video is activated to notify the older people that it is the time to take their medications.

Proteus Biomedical created a pill with embedded digestible microchip called "ingestible event markers (IEMs)" [143]. The stomach acid in the stomach is required to activate the transmission of digital signals to a microelectronic receiver. The following data can be obtained from the receiver including date, time, identify the type of medication, dosage, place of manufacture, and the changes in physiological signals as response to medication treatment, such as heart rate, respiration rate, etc. The company has claimed that capability to produce in large volumes significantly reduces the cost of sensor production. This sensor development is a fundamental part of project called 'Raisin System', which has a main goal to perform chronic diseases management such as congestive heart failure.

The SmartDrawer is designed by Becker et al. [144] to identify which bottles are taken and to recognize when the opening or closing of the drawer is initiated and to store event's timestamp in a database. There are three users of the SmartDrawer: the patient, the caregiver and the maintainer. The system provides each user with his own graphical user interface (GUI) and functionalities. The patient interface generates alarm messages when the medication is not consumed in the appropriate dosage. It also has a feature to show the patient's history of taking medication and a medication schedule including times as well as dosing information. The requirements for the interface used by the patient are must be intuitive and easy to use. The best solution for this case is a touch screen. There are two features for the caregiver view: the feature to retrieve the pattern of medication taking for each patient and the feature which is used not only to modify existing prescriptions but also to add new ones. The remote access must be provided by the system so that the number of home visit by the caregivers can be reduced and the changes can be performed remotely from the caregiver's office. The administrator of the system or the maintainer is allowed to create system modifications by adding or removing functionalities, accessing and converting the collected data to the required format. The maintainer could be a caregiver or a person with computer skills who may require to handle more complex operations. RFID technologies can be used to track the use of medication in the case where different types of medicines must be taken each day. An RFID tag embedded in each medicine bottle when taking up the bottle from the drawer, then the RFID reader will capture the identification (ID) number and the timestamp of the event. Then, data analysis will be performed to determine when there is inconsistency in medication usage. If the inconsistency exists, the SmartDrawer will notify both the patient and the caregiver.

6.4 Discussion

Three main issues arising from the adoption of wireless networks for medical application are security, privacy and the technology learning curve [11]. Providing a secure network for protecting the privacy of medical records and other health information without affecting the power of the devices is still a huge challenge for developers of medical devices. The developers also need to consider that users may find difficulties to use the device and demands a simple and helpful user manual.

7 Conclusion and Future Works

A Cyber-Physical System (CPS) with high reliability must guarantee a high level of trustworthiness [145]. The main important issue for implementing a reliable CPS is the trustworthiness of data, this situation might be partly caused by the following factors [146]:

- **Huge data size**. A typical CPS consists of hundreds of sensors, which are responsible to measure the environmental parameters. It requires an effective management system that enables the processing of large datasets, identifies abnormal situations and subsequently generates alarm signals if predefined critical levels are exceeded.
- **Noisy data**. The unreliable and noisy data are major problems faced by many researchers when developing CPS applications. A study conducted by Buonadonna et al. [147] has listed the difficulties to obtain meaningful and accurate CPS data. Most of the failures have occurred in unexpected ways and less than a half of collected data can be extracted into meaningful information. Szewczyk et al. [148] have deployed a project at Great Duck Island and has revealed that only 40 % of the data are accurate
- **Non-availability of training sets**. Many general approaches to detect false alarm are supervised learning, for example training datasets are used to build the classifiers [149]. Any data collection can be quiet costly and manually labeling a large dataset of sensor measurements could lead to error-prone work.
- **Conflicts of sensors**. Reasonable redundancies must be recognized by a well deployed CPS, for instance, deploying wireless network for k-coverage requires at least k different sensors to monitor each node directly [150]. Sometimes a sensor could be faulty, but collected measurements from the rest sensors still can be useful. There are certain cases where conflicts between reliable and faulty sensors can occur. Since the user has no idea which is the reliable one, the truth must be inferred by the system from conflicting data.
- **Uncertainty of objects**. The CPS triggers an alarm when detecting abnormal events in the monitored area, e.g., the intrusion by enemy in a battle network system. Most of the sensors can only inform the estimated area of the intrusion object but not providing the detail of the site, the reason for this are due to hardware limitation. The data integration from various sensors is required by the system to provide further details of the object.

References

1. Y. Lun, L. Cheng, The research on the model of the context-aware for reliable sensing and explanation in cyber-physical system. Procedia Eng. **15**, 1753–1757 (2011)
2. K. Wan, K. Man, D. Hughes, Specification, analyzing challenges and approaches for cyber-physical systems (cps). Eng. Lett. **18**(3), 308 (2010)
3. K.S. Nikita, J.C. Lin, D.I. Fotiandis, M.T.A. Waldmeyer, Editorial: special issue on mobile and wireless technology for healthcare delivery. IEEE Trans. Biomed. Eng. **59**(11), 3083–3089 (2012)
4. W. Swan, *Australia to 2050: future challenges*. (Commonwealth of Australia, Sydney, 2010)
5. S. Krishna, S.A. Boren, E.A. Balas, Healthcare via cell phones: a systematic review. Telemed. e-Health **15**(3), 231–240 (2009)

6. E. Kyriacou, M. Pattichis, C. Pattichis, A. Panayides, A. Pitsillides, M-health e-emergency systems: current status and future directions [wireless corner]. Antennas Propag. Mag. IEEE **49**(1), 216–231 (2007)
7. H. Alemdar, C. Ersoy, Wireless sensor networks for healthcare: a survey. Comput. Netw. **54**(15), 2688–2710 (2010)
8. E.E. Egbogah, A.O. Fapojuwo, A survey of system architecture requirements for health care-based wireless sensor networks. Sensors **11**(5), 4875–4898 (2011)
9. S. Ullah, P. Khan, N. Ullah, S. Saleem, H. Higgins, K.S. Kwak, A review of wireless body area networks for medical applications. Int. J. Commun. Netw. Syst. Sci. **2**, 797–803 (2009)
10. A. Wood, G. Virone, T. Doan, Q. Cao, L. Selavo, Y. Wu, L. Fang, Z. He, S. Lin, J. Stankovic, Alarm-net: wireless sensor networks for assisted-living and residential monitoring. Technical Report, Department of Computer Science, University of Virginia, 2006
11. G. Shobha, R.R. Chittal, K. Kumar, Medical applications of wireless networks, in *Proceedings of the 2nd International Conference on Systems and Networks Communications*, 2007, pp. 82–82
12. E. Jovanov, Wireless technology and system integration in body area networks for m-health applications, in *Proceedings of the 27th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2006, pp. 7158–7160
13. H. Huo, Y. Xu, H. Yan, S. Mubeen, H. Zhang, An elderly health care system using wireless sensor networks at home, in *Proceedings of the 3rd International Conference on Sensor Technologies and Applications*, 2009, pp. 158–163
14. V. Hessels, G.S. Le Prell, W.C. Mann, Advances in personal emergency response and detection systems. Assistive Technol. **23**(3), 152–161 (2011)
15. W.C. Mann, P. Belchior, M.R. Tomita, B.J. Kemp, Use of personal emergency response systems by older individuals with disabilities. Assistive Technol. **17**(1), 82–88 (2005)
16. E. Porter, Wearing and using personal emergency response system buttons: older frail widows' intentions. J. Gerontol. Nurs. **31**, 26–33 (2005)
17. M. Bernstein, "Low-tech" personal emergency response systems reduce costs and improve outcomes. Manag. Care Q. **8**(1), 38 (2000)
18. D.A. Levine, R. Tideiksaar et al., Personal emergency response systems: factors associated with use among older persons. Mount Sinai J. Med. **62**(4), 293 (1995)
19. M. Hamill, V. Young, J. Boger, A. Mihailidis, Development of an automated speech recognition interface for personal emergency response systems. J. NeuroEng. Rehabil. **6**(1), 26 (2009)
20. J. Lau, Building a national technology and innovation infrastructure for an aging society, Ph.D. Dissertation, University of Pennsylvania, 2005
21. R.W. Pew, S.B. Van Hemel et al., *Technology for Adaptive Aging*. (National Academies Press, Washington DC, 2004)
22. M.J.S. Gibson, R.O. Andres, B. Isaacs, T. Radebaugh, J. Worm-Petersen, The prevention of falls in later life. Dan. Med. Bull. **34**(4), 1–24 (1987)
23. A. Tovell, K. McKenna, C. Bradley, S. Pointer, Hospital separations due to injury and poisoning, Australia, Technical Report, Australian Institute of Health and Welfare, 2012
24. C. Bradley, Hospitalisations due to falls in older people, Australia, 2003–2004, Technical Report, Australian Institute of Health and Welfare, 2012
25. K. Hauer, S.E. Lamb, E.C. Jorstad, C. Todd, C. Becker et al., Systematic review of definitions and methods of measuring falls in randomised controlled fall prevention trials. Age Ageing **35**(1), 5–10 (2006)
26. X. Yu, Approaches and principles of fall detection for elderly and patient, in *Proceedings of the 10th International Conference on e-health Networking, Applications and Services*, 2008, pp. 42–47
27. T. Lee, A. Mihailidis, An intelligent emergency response system: preliminary development and testing of automated fall detection. J. Telemed. Telecare **11**(4), 194–198 (2005)
28. A. Ariani, Simulation of a wireless sensor network for unobtrusively detecting falls in the home, Ph.D. Dissertation, University of New South Wales, 2012

29. A. Leone, G. Diraco, P. Siciliano, Detecting falls with 3d range camera in ambient assisted living applications: a preliminary study. Med. Eng. Phys. **33**(6), 770–781 (2011)
30. J. Shim, M.-h. Shim, Y.-s. Baek, T.-d. Han, The development of a detection system for seniors' accidental fall from bed using cameras, in *Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication*, 2011, p. 102
31. Y. Zigel, D. Litvak, I. Gannot, A method for automatic fall detection of elderly people using floor vibrations and sound-proof of concept on human mimicking doll falls, IEEE T. Bio-Med. Eng. **56**(12), 2858–2867 (2009)
32. J. Willems, G. Debard, B. Vanrumste, T. Goedeḿe, A video-based algorithm for elderly fall detection, in world congress on medical physics and biomedical engineering, Munich, Germany, 7–12 Sept, 2009, pp. 312–315
33. R. Steele, A. Lo, C. Secombe, Y.K. Wong, Elderly persons' perception and acceptance of using wireless sensor networks to assist healthcare. Int. J. Med. Inform. **78**(12), 788–801 (2009)
34. A. Rowe, A resource-centric design paradigm for scalable sensor networks, Ph.D. Dissertation, Carnegie Mellon University, 2010
35. Y. Zigel, D. Litvak, I. Gannot, A method for automatic fall detection of elderly people using floor vibrations and sound-proof of concept on human mimicking doll falls. IEEE Trans. Biomed. Eng. **56**(12), 2858–2867 (2009)
36. O. Ojetola, E.I. Gaura, J. Brusey, Fall detection with wearable sensors–safe (smart fall detection), in *Proceedings of the 7th International Conference on Intelligent Environments*, 2011, pp. 318–321
37. M. Tolkiehn, L. Atallah, B. Lo, G.-Z. Yang, Direction sensitive fall detection using a triaxial accelerometer and a barometric pressure sensor, in *Proceedings of the 32th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2011, pp. 369–372
38. V.Q. Viet, G. Lee, D. Choi, Fall detection based on movement and smart phone technology, in *Proceedings of the 2012 IEEE RIVF International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF)*, 2012, pp. 1–4
39. M. Alwan, P.J. Rajendran, S. Kell, D. Mack, S. Dalal, M. Wolfe, R. Felder, A smart and passive floor-vibration based fall detector for elderly, in *Proceedings of the 2nd International Conference on Information & Communication Technologies: from Theory to Applications*, 2006, pp. 1003–1007
40. M. Stikic, T. Huynh, K.V. Laerhoven, B. Schiele, ADL recognition based on the combination of RFID and accelerometer sensing, in *Proceedings of the 2nd International Conference on Pervasive Computing Technologies for Healthcare*, 2008, pp. 258–263
41. D.L. Algase, Biomechanical activity devices to index wandering behaviour in dementia. Am. J. Alzheimer Dis. Dement. **18**(2), 85–92 (2003)
42. A. Ariani, S.J. Redmond, D. Chang, N.H. Lovell, Software simulation of unobtrusive falls detection at night-time using passive infrared and pressure mat sensors, in *Proceedings of the 32rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2010, pp. 2115–2118
43. D. Litvak, I. Gannot, Y. Zigel, Detection of falls at home using floor vibrations and sound, in *Proceedings of the IEEE 25th Convention of Electrical and Electronics Engineers*, 2008, pp. 514–518
44. T. Liu, X. Guo, G. Wang, Elderly-falling detection using distributed direction-sensitive pyroelectric infrared sensor arrays. Multidimens. Syst. Signal Process. **23**(4), 451–467 (2012)
45. A. Ariani, S.J. Redmond, D. Chang, N.H. Lovell, Simulated unobtrusive falls detection with multiple persons. IEEE Trans. Biomed. Eng. **59**(11), 3185–3196 (2012)
46. A.R. Kaushik, B.G. Celler, Characterization of passive infrared sensors for monitoring occupancy pattern, in *Proceedings of the 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2006, pp. 5257–5260

47. N. Cumming, *Security: A Guide to Security System Design and Equipment Selection and Installation*. (Butterworth-Heinemann, Oxford, 1994), ch.4, pp. 115–176
48. N. Noury, A. Galay, J. Pasquier, M. Ballussaud, Preliminary investigation into the use of autonomous fall detectors, in *Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2008, pp. 2828–2831
49. K. Johnston, A. Worley, K. Grimmer-Somers, M. Sutherland, L. Amos, Personal alarm use to call the ambulance after a fall in older people: characteristics of clients and falls. J. Emerg. Prim. Health Care **8**(4), 1–9 (2010)
50. C.A. Otto, X. Chen, Automated fall detection: saving senior lives one fall at a time. *Caring: National Association for Home Care magazine*. **28**(3) 44 (2009)
51. G. Diraco, A. Leone, P. Siciliano., An active vision system for fall detection and posture recognition in elderly healthcare, in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2010, pp. 1536–1541
52. S. McLean, D. Protti, A. Sheikh, Telehealthcare for long term conditions. BMJ **342**, 374–378 (2011)
53. E.B. Allely, Synchronous and asynchronous telemedicine. J. Med. Syst. **19**(3), 207–212 (1995)
54. S.D. Anker, F. Koehler, W.T. Abraham, Telemedicine and remote management of patients with heart failure. Lancet **378**, 731–739 (2011)
55. M. Raad, L. Yang, A ubiquitous smart home for elderly, Inf. Syst. Front. **11(**5) 1–5 (2009)
56. U. Hansmann, L. Merk, M.S. Nicklous, T. Stober, *Pervasive Computing: The mobile world* (Springer, New York, 2003)
57. J. Kim, M. Kang, B. Hwang, A method for detecting arrhythmia using a RR interval from ECG data in u-health system, in *Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication*, 2011, p. 15
58. B. Baby, M.S. Manikandan, K. Soman, Automated cardiac event change detection for continuous remote patient monitoring devices, in *Proceedings of the 1st International Conference on Wireless Technologies for Humanitarian Relief*, 2011, pp. 225–232
59. K. Li, N. Du, A. Zhang, Detecting ECG abnormalities via transductive transfer learning, in *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, 2012, pp. 210–217
60. V. Gay, P. Leijdekkers, E. Barin, A mobile rehabilitation application for the remote monitoring of cardiac patients after a heart attack or a coronary bypass surgery, in *Proceedings of the 2nd International Conference on Pervasive Technologies Related to Assistive Environments*, 2009, p. 21
61. H. Witte, L.D. Iasemidis, B. Litt, Special issue on epileptic seizure prediction. IEEE Trans. Biomed. Eng. **50**(5), 537–539 (2003)
62. A. Dalton, S. Patel, A. Chowdhury, M. Welsh, T. Pang, S. Schachter, G. OLaighin, P. Bonato, Development of a body sensor network to detect motor patterns of epileptic seizures. IEEE Trans. Biomed. Eng. **59**(11), 3204–3211 (2012)
63. C. Deckers, P. Genton, G. Sills, D. Schmidt et al., Current limitations of antiepileptic drug therapy: a conference review. Epilepsy Res. **53**(1–2), 1 (2003)
64. W.O. Tatum IV, L. Winters, M. Gieron, E.A. Passaro, S. Benbadis, J. Ferreira, J. Liporace, Outpatient seizure identification: results of 502 patients using computer-assisted ambulatory EEG. J. Clin. Neurophysiol. **18**(1), 14–19 (2001)
65. D. Schmidt, C. Elger, G.L. Holmes, Pharmacological overtreatment in epilepsy: mechanisms and management. Epilepsy Res. **52**(1), 3–14 (2002)
66. A.H. Shoeb, J.V. Guttag, Application of machine learning to epileptic seizure detection, in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 975–982
67. A. Economics, *Listen Hear!: The Economic Impact and Cost of Hearing Loss in Australia* (Access Economics, Sydney, 2006)
68. P. Hyvärinen, Utilization of the chirp stimulus in auditory brainstem response measurements, Ph.D. Dissertation, Aalto University, 2012

69. W.G. Noble, *Assessment of Hearing Impairment: A Critique and a New Method* (Academic Press, New York, 1978)
70. D.S. Dalton, K.J. Cruickshanks, B.E. Klein, R. Klein, T.L. Wiley, D.M. Nondahl, The impact of hearing loss on quality of life in older adults. Gerontologist **43**(5), 661–668 (2003)
71. A. Al-Afsaa, S. Soegijoko, Development of a chirp stimulus pc-based auditory brainstem response audiometer. ITB J. Eng. Sci. **36**(1), 81–94 (2004)
72. T. Dau, O. Wegner, V. Mellert, B. Kollmeier, Auditory brainstem responses with optimized chirp signals compensating basilar-membrane dispersion. J. Acoustical Soc. Am. **107**, 1530–1540 (2000)
73. K.R. Taylor, A.P. DeLuca, A. Eliot Shearer, M.S. Hildebrand, E. Ann Black-Ziegelbein, V. Nikhil Anand, C. M. Sloan, R.W. Eppsteiner, T.E. Scheetz, P.L. Huygen et al., Audiogene: predicting hearing loss genotypes from phenotypes to guide genetic screening, Hum. Mutat. **34**(4) 539–545 (2012)
74. ABS, *Musculoskeletal Conditions in Australia: A Snapshot, 2004–05* (Australian Bureau of Statistic, Canberra, 2006)
75. ABS, *Injury in Australia: A snapshot, 2004-05* (Australian Bureau of Statistic, Canberra, 2006)
76. P. Hattam, The effectiveness of orthopaedic triage by extended scope physiotherapists. Clin. Gov. Int. J. **9**(4), 244–252 (2004)
77. T. Russell, P. Truter, R. Blumke, B. Richardson, The diagnostic accuracy of telerehabilitation for nonarticular lower-limb musculoskeletal disorders. Telemed. e-Health **16**(5), 585–594 (2010)
78. O.S. Lowe, Australia: services for Australian rural and remote allied health, inc. National Allied Health Workforce Report, Technical Report, 2004
79. D. Theodoros, T. Russell et al., Telerehabilitation: current perspectives. Stud. Health Technol. Inform. **131**, 191–210 (2008)
80. Z. Cao, R. Zhu, R. Que, A wireless portable system with micro sensors for monitoring respiratory diseases. IEEE Trans. Biomed. Eng. **59**(11), 3110–3116 (2012)
81. J.G. Park, K. Ramar, E.J. Olson, Updates on definition, consequences, and management of obstructive sleep apnea. Mayo Clin. Proc. **86**(6), 549–555 (2011)
82. W. Lee, S. Nagubadi, M.H. Kryger, B. Mokhlesi, *Epidemiology of Obstructive Sleep Apnea: A Population-Based Perspective* (Expert Reviews Ltd, London, 2008)
83. T. Young, P.E. Peppard, D.J. Gottlieb, Epidemiology of obstructive sleep apnea a population health perspective. Am. J. Respir. Crit. Care Med. **165**(9), 1217–1239 (2002)
84. C.D. Mathers, D. Loncar, Projections of global mortality and burden of disease from 2002 to 2030. PLoS Med. **3**(11), e442 (2006)
85. F. Agostinis, C. Foglia, M. Landi, M. Cottini, C. Lombardi, G.W. Canonica, G. Passalacqua, Gina report, global strategy for asthma management and prevention. Allergy **63**(12), 1637–1639 (2008)
86. M.J. Coma-del Corral, M.L. Alonso-Álvarez, M. Allende, J. Cordero, E. Ordax, F. Masa, J. Terán-Santos, Reliability of telemedicine in the diagnosis and treatment of sleep apnea syndrome. Telemed. e-Health **19**(1) 7–12 (2013)
87. R.M. Sapolsky, G. Gurley, D. Demarest, E. internationales Télé-Film, *Why zebras Don't Get Ulcers* (Times Books, New York, 2004)
88. F.H. Wilhelm, P. Grossman, Emotions beyond the laboratory: theoretical fundaments, study design, and analytic strategies for advanced ambulatory assessment. Biol. Psychol. **84**(3), 552–569 (2010)
89. E. Ertin, N. Stohs, S. Kumar, A. Raij, M. al'Absi, S. Shah, Autosense: unobtrusively wearable sensor suite for inferring the onset, causality, and consequences of stress in the field, in *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*, 2011, pp. 274–287
90. K.D. Kochanek, J. Xu, S.L. Murphy, A.M. Minino, H.-C. Kung, Deaths: preliminary data for 2009. Nat. Vital Stat. Rep. **59**(4), 1–51 (2011)

91. S.L. Murphy, J. Xu, K.D. Kochanek, Deaths: preliminary data for 2010. Nat. Vital Stat. Rep. **60**(4), 1–51 (2012)
92. V.L. Roger, A.S. Go, D.M. Lloyd-Jones, E.J. Benjamin, J.D. Berry, W.B. Borden, D.M. Bravata, S. Dai, E.S. Ford, C.S. Fox et al., Heart disease and stroke statistics 2012 update a report from the American heart association. Circulation **125**(1), e2–e220 (2012)
93. L.B. Goldstein, C.D. Bushnell, R.J. Adams, L.J. Appel, L.T. Braun, S. Chaturvedi, M.A. Creager, A. Culebras, R.H. Eckel, R.G. Hart et al., Guidelines for the primary prevention of stroke a guideline for healthcare professionals from the American heart association/ American stroke association. Stroke **42**(2), 517–584 (2011)
94. J.N. Brownstein, Addressing heart disease and stroke prevention through comprehensive population-level approaches. Prev. Chronic Dis. **5**(2), A31 (2008)
95. X. Ma, X. Tu, J. Huang, J. He, A cyber-physical system based framework for motor rehabilitation after stroke, in *Proceedings of the 1st International Conference on Wireless Technologies for Humanitarian Relief*, 2011, pp. 285–290
96. D.J. Foley, A.A. Monjan, S.L. Brown, E.M. Simonsick et al., Sleep complaints among elderly persons: an epidemiologic study of three communities. Sleep J. Sleep Res. Sleep Med. **18(**6) 425–432 (1995)
97. D. Foley, S. Ancoli-Israel, P. Britz, J. Walsh, Sleep disturbances and chronic disease in older adults: results of the 2003 national sleep foundation sleep in america survey. J. Psychosom. Res. **56**(5), 497–502 (2004)
98. H. Ni, B. Abdulrazak, D. Zhang, S. Wu, Unobtrusive sleep posture detection for eldercare in smart home, in *Aging Friendly Technology for Health and Independence* (Springer, New York, 2010), pp. 67–75
99. WHO, *Global Tuberculosis Report 2012* (World Health Organization, Geneva, 2012)
100. A.P. Koesoema, Y.S. Irawan, S. Soegijoko, Preliminary design of a community telemedicine system for tuberculosis control, in *World Congress on Medical Physics and Biomedical Engineering 2006*, 2007, pp. 366–369
101. H. Rachmat, L.I. Octavia, S. Soegijoko, Development of a simple e-health system for tuberculosis management at community health center level in Indonesia, in *Med-e-Tel 2009 Proceedings*, 2009, pp. 366–369
102. S. Soegijoko, Application specific e-health & telemedicine systems: implementation experience for community healthcare and systematic review of disaster publications, in *Proceedings of the 2nd International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering*, 2011, pp. 415–416
103. E. Priya, S. Srinivasan, Automated decision support system for tuberculosis digital images using evolutionary learning machines. Eur. J. Biomed. Inform. **9** en3–en8 (2013)
104. A.U. Alahakone, S.A. Senanayake, A real-time system with assistive feedback for postural control in rehabilitation. IEEE/ASME Trans. Mechatron. **15**(2), 226–233 (2010)
105. O. Aziz, L. Atallah, B. Lo, M. ElHelw, L. Wang, G.-Z. Yang, A. Darzi, A pervasive body sensor network for measuring postoperative recovery at home. Surg. Innovation **14**(2), 83–90 (2007)
106. E. Monton, J. Hernandez, J. Blasco, T. Herve, J. Micallef, I. Grech, A. Brincat, V. Traver, Body area network for wireless patient monitoring. IET Commun. **2**(2), 215–222 (2008)
107. V. Leonov, T. Torfs, P. Fiorini, C. Van Hoof, Thermoelectric converters of human warmth for self-powered wireless sensor nodes. Sens. J. IEEE **7**(5), 650–657 (2007)
108. B.-S. Lin, B.-S. Lin, N.-K. Chou, F.-C. Chong, S.-J. Chen, RTWPMS: a real-time wireless physiological monitoring system. IEEE Trans. Inf Technol. Biomed. **10**(4), 647–656 (2006)
109. A. Milenkovi′c, C. Otto, E. Jovanov, Wireless sensor networks for personal health monitoring: Issues and an implementation. Comput. Commun. **29**(13), 2521–2533 (2006)
110. J. Yao, R. Schmitz, S. Warren, A wearable point-of-care system for home use that incorporates plug-and-play and wireless standards. IEEE Trans. Inf Technol. Biomed. **9**(3), 363–371 (2005)
111. Z. Lv, F. Xia, G. Wu, L. Yao, Z. Chen, iCare: a mobile health monitoring system for the elderly, in *Proceedings of the 2010 IEEE/ACM Int'l Conference on Green Computing and*

Communications & Int'l Conference on Cyber, Physical and Social Computing, 2010, pp. 699–705
112. W. Wu, J. Cao, Y. Zheng, Y.-P. Zheng, Waiter: a wearable personal healthcare and emergency aid system, in *Proceedings of the 6th Annual IEEE International Conference on Pervasive Computing and Communications*, 2008, pp. 680–685
113. A. Pantelopoulos, N.G. Bourbakis et al., Prognosis-a wearable health-monitoring system for people at risk: methodology and modeling. IEEE Trans. Inf Technol. Biomed. **14**(3), 613–621 (2010)
114. H.-T. Chang, C.-G. Chung, M.-W. Chen, An e-caring chair for physiological signal measurement and recording. Med. Eng. Phys. **35**, 277–282 (2013)
115. G.F. Fletcher, G. Balady, S.N. Blair, J. Blumenthal, C. Caspersen, B. Chaitman, S. Epstein, E.S.S. Froelicher, V.F. Froelicher, I.L. Pina et al., Statement on exercise: Benefits and recommendations for physical activity programs for all Americans a statement for health professionals by the committee on exercise and cardiac rehabilitation of the council on clinical cardiology, American heart association. Circulation **94**(4), 857–862 (1996)
116. L. Mo, S. Liu, R. Gao, D. John, J. Staudenmayer, P. Freedson, Wireless design of a multi-sensor system for physical activity monitoring. IEEE Trans. Biomed. Eng. **59**(11), 3230–3237 (2012)
117. A. Aridarma, T. Mengko, S. Soegijoko, Personal medical assistant: Future exploration, in *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics (ICEEI)*, 2011, pp. 1–6
118. E. Sutjiredjeki, S. Soegijoko, T.L. R. Mengko, S. Tjondronegoro, K. Astami, H. U. Muhammad et al., Application of a mobile telemedicine system with multi communication links for disaster reliefs in indonesia, in *World Congress on Medical Physics and Biomedical Engineering 2009*, 2009, pp. 344–347
119. E. Sutjiredjeki, S. Soegijoko, Development of a communication arbiter for mobile telemedicine system with multi communication links, in *World Congress on Medical Physics and Biomedical Engineering 2006*, 2007, pp. 715–718
120. E. Sutjiredjeki, S. Soegijoko, T.L.R. Mengko, S. Tjondronegoro, Development of a mobile telemedicine system with multi communication links for urban and rural areas in Indonesia, in *Proceedings of the 3rd Kuala Lumpur International Conference on Biomedical Engineering*, 2007, pp. 660–663
121. A. Gaddam, S. Mukhopadhyay, G. Sen Gupta, H. Guesgen, Wireless sensors networks based monitoring: Review, challenges and implementation issues, in *Proceedings of the 3rd International Conference on Sensing Technology*, 2008, pp. 533–538
122. J.-C. Huang, Innovative health care delivery system: a questionnaire survey to evaluate the influence of behavioral factors on individuals' acceptance of telecare. Comput. Biol. Med. **43**, 281–286 (2013)
123. M. Souil and A. Bouabdallah, On QoS provisioning in context-aware wireless sensor networks for healthcare, in *Proceedings of the 20th International Conference on Computer Communications and Networks*, 2011, pp. 1–6
124. A. Mihailidis, A. Cockburn, C. Longley, J. Boger, The acceptability of home monitoring technology among community-dwelling older adults and baby boomers. Assistive Technol. **20**(1), 1–12 (2008)
125. D. Malan, T. Fulford-Jones, M. Welsh, S. Moulton, Codeblue: an ad hoc sensor network infrastructure for emergency medical care, in *Proceedings of the MobiSys 2004 Workshop on Applications of Mobile Embedded Systems*, 2004, pp. 12–14
126. M. Sung, C. Marci, A. Pentland, Wearable feedback systems for rehabilitation. J. Neuroeng. Rehabil. **2**(17), 1–12 (2005)
127. S. Brage, N. Brage, P. Franks, U. Ekelund, N. Wareham, Reliability and validity of the combined heart rate and movement sensor actiheart. Eur. J. Clin. Nutr. **59**(4), 561–570 (2005)

128. T. Huynh, B. Schiele, Analyzing features for activity recognition, in *Proceedings of the 2005 Joint Conference on Smart Objects and Ambient Intelligence: Innovative Context Aware Services: Usages and Technologies*, 2005, pp. 159–163
129. P. Lukowicz, J.A. Ward, H. Junker, M. Stäger, G. Tröster, A. Atrash, T. Starner, Recognizing workshop activity using body worn microphones and accelerometers, in *Pervasive Computing* (Springer, 2004), pp. 18–32
130. D.L. Hall, J. Llinas, An introduction to multisensor data fusion. Proc. IEEE **85**(1), 6–23 (1997)
131. D. Kotz, S. Avancha, A. Baxi, A privacy framework for mobile health and home-care systems, in *Proceedings of the 1st ACM workshop on Security and privacy in medical and home-care systems*, 2009, pp. 1–12
132. R. Simpson, D. Schreckenghost, E.F. LoPresti, N. Kirsch, Plans and planning in smart homes, in *Designing Smart Homes* (Springer, New York, 2006), pp. 71–84
133. G. Mantas, D. Lymberopoulos, N. Komninos, Integrity mechanism for e-health telemonitoring system in smart home environment, in *Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2009, pp. 3509–3512
134. M. Guennoun, K. El-Khatib, Securing medical data in smart homes, in *Proceedings of the 4th IEEE International Workshop on Medical Measurements and Applications*, 2009, pp. 104–107
135. B. Fong, M. Pecht, Prognostics in wireless telecare networks: a perspective on serving the rural chinese population, in *Proceedings of the 2010 Annual Conference of the Prognostics and Health Management Society*, 2010, pp. 1–6
136. M. Wang, K. Fishkin, *A Flexible, Low-Overhead Ubiquitous System for Medication Monitoring* (Intel Corporation, Tech. Rep., 2003)
137. J.K. Vinjumur, E. Becker, S. Ferdous, G. Galatas, F. Makedon, Web based medicine intake tracking application, in *Proceedings of the 3rd International Conference on Pervasive Technologies Related To Assistive Environments*. ACM, 2010, p. 37
138. D. Hess, R.M. Kacmarek, M.H. Kollef, *Essentials of Mechanical Ventilation* (McGraw-Hill, Health Professions Division, New York, 1996)
139. A.M. Cheng, Cyber-physical medical and medication systems, in *Proceedings of the 28th International Conference on Distributed Computing Systems Workshop*, 2008, pp. 529–532
140. M. Pajic, R. Mangharam, O. Sokolsky, D. Arney, J. Goldman, I. Lee, Model-driven safety analysis of closed-loop medical systems. IEEE Trans. Industr. Inf. **99**, 1–13 (2012)
141. T.K. Nuckols, A.G. Bower, S.M. Paddock, L.H. Hilborne, P. Wallace, J.M. Rothschild, A. Griffin, R.J. Fairbanks, B. Carlson, R.J. Panzer, R.H. Brook, Programmable infusion pumps in icus: an analysis of corresponding adverse drug events. J. Gen. Intern. Med. **23**(1), 41–45 (2008)
142. J.T. Matthews, Existing and emerging healthcare devices for elders to use at home. Generations **30**(2), 13–19 (2006)
143. K. Daniel, C.L. Cason, S. Ferrell, Assistive technologies for use in the home to prolong independence, in *Proceedings of the 2nd International Conference on PErvasive Technologies Related to Assistive Environments*, 2009, p. 26
144. E. Becker, V. Metsis, R. Arora, J. Vinjumur, Y. Xu, F. Makedon, Smartdrawer: RFID based smart medicine drawer for assistive environments, in *Proceedings of the 2nd International Conference on Pervasive Technologies Related to Assistive Environments*, 2009, p. 49
145. L.-A. Tang, X. Yu, S. Kim, Q. Gu, J. Han, A. Leung, T. La Porta, Trustworthiness analysis of sensor data in cyber-physical systems. J. Comput. Syst. Sci. **79**, 383–401 (2012)
146. E.A. Lee, Cyber physical systems: design challenges, in *Proceedings of the 11th IEEE International Symposium on Object Oriented Real-Time Distributed Computing (ISORC)*, 2008, pp. 363–369
147. P. Buonadonna, D. Gay, J.M. Hellerstein, W. Hong, S. Madden, Task: sensor network in a box, in *Proceedings of the 2nd European Workshop on Wireless Sensor Networks*, 2005, pp. 133–144

148. R. Szewczyk, J. Polastre, A. Mainwaring, D. Culler, Lessons from a sensor network expedition, in *Wireless Sensor Networks*. Springer, 2004, pp. 307–322
149. K. Ni, G. Pottie, Bayesian selection of non-faulty sensors, in *Proceedings of the IEEE International Symposium on Information Theory*, 2007, pp. 616–620
150. H. Gupta, Z. Zhou, S.R. Das, Q. Gu, Connected sensor cover: self-organization of sensor networks for efficient query execution. IEEE/ACM Trans. Netw. **14**(1), 55–67 (2006)

Context-Aware Perception for Cyber-Physical Systems

Shaheena Noor and Humera Noor Minhas

Abstract Being aware of the context is one the important requirements of Cyber-Physical Systems (CPS). Context-aware systems have the capability to sense what is happening or changing in their environment and take appropriate actions to adapt to the changes. In this chapter, we present a technique for identifying the focus of attention in a context-aware cyber-physical system. We propose to use first-person vision, obtained through wearable gaze-directed camera that can capture the scene through the wearer's point-of-view. We use the fact that human cognition is linked to his gaze and typically the object/person of interest holds our gaze. We argue that our technique is robust and works well in the presence of noise and other distracting signals, where the conventional techniques of IR sensors and tagging fail. Moreover, the technique is unobtrusive and does not pollute the environment with unnecessary signals. Our approach is general in that it may be applied to a generic CPS like healthcare, office and industrial scenarios and also in intelligent homes.

Keywords First person vision · Gaze-directed vision · Context-aware perception · Cognition-action linkage · Video-based object recognition

1 Introduction

The advancements in mobile, wireless and embedded technologies, miniaturization of computing devices and reduction in cost of electronic equipment in recent years has led its way for Cyber Physical Systems (CPS) to enter our day-to-day

S. Noor (✉)
Department of Computer Engineering, Sir Syed University of Engineering and Technology, Karachi, Pakistan
e-mail: shanoor@ssuet.edu.pk

H. N. Minhas
Project Coordinator and Academic Supervisor, German Academic Exchange Service (DAAD), University of Bremen (UB), Bremen, Germany
e-mail: humera.noor@gmail.com

Z. H. Khan et al. (eds.), *Computational Intelligence for Decision Support in Cyber-Physical Systems*, Studies in Computational Intelligence 540,
DOI: 10.1007/978-981-4585-36-1_4,

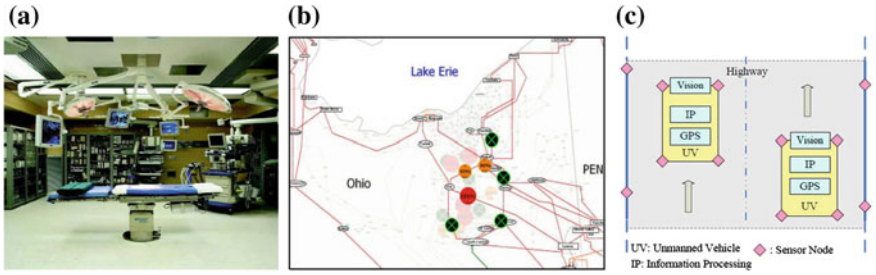


Fig. 1 Sample cyber physical systems: **a** An operating room [1, 2] **b** Electric power grid [1, 2] **c** Intelligent road with unmanned vehicle [3]

lives. CPS—characterized by integrating computation and physical processes—have applications in domains like communication, energy, aerospace, automotive, civil infrastructure, healthcare, military, robotics and transportation etc. Consider Fig. 1 [1–3] that shows three sample CPS in real-world.

Being aware of the context is one the important requirements of CPS. Context-aware systems have the capability to sense what is happening or changing in their environment and take appropriate actions to adapt to the changes. Lack of context-awareness has been shown (e.g. in [4]) to lead to interruptions resulting in increased task difficulty, completion time, errors and annoyance while reducing the overall working efficiency. Three important aspects of context are to identify where you are; who you are with; and which resources are nearby. Things become complicated when it is required to identify the exact and current focus of attention in a scenario with multiple elements of interest.

The often ignored element in the loop—the human—plays a significant role in the CPS setup and has recently started to receive attention. In the Human-in-the-Loop CPS (HiLCPS) systems, the human cognitive activity is noted and analyzed to interpret his intent, which serves as an input to the autonomous device's control system (e.g. an assistive robot) [5]. In such scenarios, the human serves as a major guide to find the missing pieces of the jigsaw puzzle and provides the much-needed context-awareness of the environment. To this end, vision is considered as one of the most important cues for perception and action, and much of object, scene and environment perception is based on visual input.

Traditionally, noting human behavior and activities involved camera(s) mounted at fixed locations in the scene and subjects were asked to perform a task including navigation, manipulation etc. This may be regarded as outside-in vision, where the analysis is carried from the notion of a third person observing someone doing an activity. However, the outside-in view is sometimes neither sufficient nor possible; hence a first-person perspective becomes desirable. The study of human working behavior reveals the importance of eye gaze in performing any task at hand. Human gaze and eye tracking offer insight into the cognitive processes involved in the tasks being carried out [6–8]. First Person Vision, aka Inside-Out vision is obtained via a wearable, eye-controlled camera that is always directed to

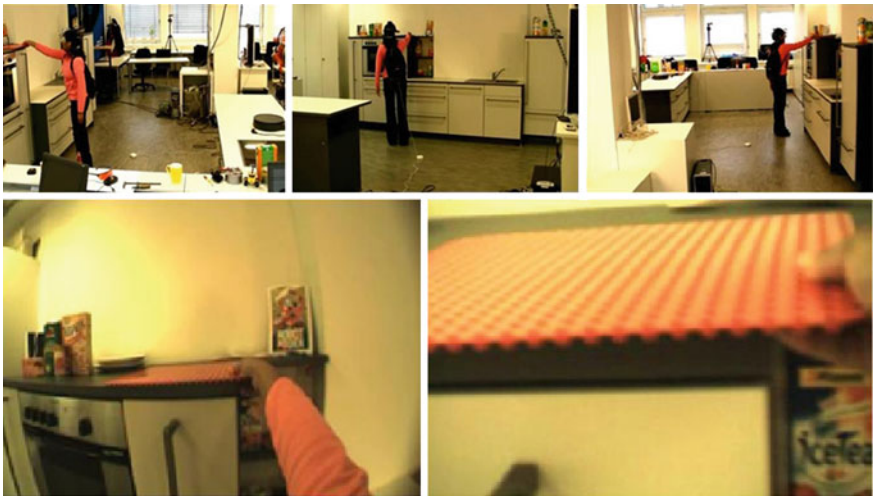


Fig. 2 Inside-out and outside-in views of a scene

the target of gaze. The vision-cognition linkage has given inside-out vision tremendous importance and allowed to view the environment from ego-centric perspective. Consider Fig. 2 that shows the first and third person perspectives of a person performing some activity in a scene. While outside-in view (top) provides overall information of the environment including actor's pose and activity information, the inside-out vision (bottom) provides the exact attention information. Together, the two views provide the context-awareness, which is desirable in CPS.

In this chapter, we present a robust mechanism to identify the focus of attention of the personnel in a CPS by considering a pervasive health care scenario. We propose to use wearable gaze-directed cameras that can capture the exact point where the person is looking to identify his focus. We exploit the fact that it is human nature to "look" at the relevant object and/or person attended to and human gaze can point to his focus of attention and thought process. Using the same setup, we provide location-awareness to the accuracy of exact room and/or corridor in a huge, complex building just by having the person look around the place where he's present. Our system also allows browsing through personnel history and different tabs on screen just by looking at the appropriate tab on a handheld device.

2 Background Review

The foundation of Cyber Physical Systems is considered to be laid as early as the development of microprocessors in the era of 1970s [9]. Ever since the technology is advancing and CPS are now being visualized as the next generation of computing. Much work is being done on one hand to employ them into real-world

application e.g. [1–3] and on the other hand to identify their core weaknesses, the way-arounds and presentation of strong infrastructure e.g. [10–13].

Lee in [10] argues that there lies a huge gap between the underlying computing foundations between CPS and the real world, which impedes technical progress. He supports the use of real-time OS, middleware technologies, and specialized processing and networking devices to bridge the gap. Further challenges, as pointed out by Sha et al. in [11], are real-time system abstractions, robustness and safety issues, system QoS composition and system trustworthiness.

One of the major application areas of CPS is medical and healthcare systems because of the criticality and sensitivity of the domain. According to [14], in a study conducted in USA, thousands of patients die every year in the hospitals due to medical errors, most of which are actually avoidable. A context-aware pervasive health care system can help in minimizing the loopholes and points of failures and provide procedural improvements and ameliorate monitoring and diagnostics resulting in enhanced quality of patient care, increased staff efficiency and reduced capital and operational costs. Sha et al. in [11] and Shi et al. in [3] present Medical Network Device as a sample CPS to highlight the new frontiers explored and benefits obtained by introducing such a setup.

Since deploying a context-aware healthcare solution benefits the entire hospital, including business decision makers, technical decision makers, clinical staff and above-all the patients, much effort is being put into make the setup more and more secure in terms of confidentiality, integrity, availability and accountability [15]. In [16], Bardram et al. present an activity-driven computing infrastructure focusing on preserving user's context in the event of interruption and mobility. On the other hand, Zhang et al. in [17] focused on having an infrastructure capable of personalization by identifying the logical components. Jahnke et al. in [18] proposes to have context-awareness as implicit as possible so as to avoid any resistance by the clinicians and patients towards the electronic technology. A survey on context-aware systems was done by Baldauf et al. in [19], wherein they presented the common architecture principles of context-aware systems with special emphasis on context-aware middleware and frameworks which ease in the application development. Orwat et al. presented another review in [20] on the developments and implementations of pervasive computing systems in healthcare. More surveys on pervasive healthcare systems were conducted by Bricon-Souf and Newman [21] and Miraoui et al. [22].

In order to provide a seamless transmission of information, the users (e.g. patients, doctors) and objects (e.g. equipment) are being equipped with wearable devices like location sensor, identification tags and even devices to note medical parameters like temperature, blood pressure etc [23]. As the objects, actors and their communications in a context-aware environment increases, and the need to involve human beings in the loop, so does the problem of identifying the focus of attention at any given point in time. Hence, much work concentrates specifically on attention-awareness and attention-aware systems e.g. [24, 25]. The study of human working behavior reveals the importance of eye gaze in performing any task at hand. Human gaze and eye tracking offer insight into the cognitive

processes involved in the tasks being carried out [7]. This was primarily revealed in the classical experiments by Yarbus [8], which formed the basis of further exploration [6]. Eye gaze fixations and motions are strongly linked to the cognitive and thought process which in turn is related to the goal/objective that we're trying to reach. In this connection visual cues are considered as a vital enabling mechanism for identifying focus of attention e.g. by Toet in [26]. Ever since the relation between human cognitive process and eye movements has been discovered, many attempts have been made to track and monitor human gaze e.g. [27–31]. As a result further insight into this relationship is obtained. In the daily routine tasks like household chores of making a sandwich/tea or driving a car the spatio-temporal demands of the job play a significant role [6, 29, 32]. Humans tend to develop a knowledge-base, by gazing around aimlessly, as soon as they enter an environment [29, 32]. Such knowledge-base helps in the eventual execution of tasks as it gives a rough estimation of objects in the scene and their locations. A number of experiments have been carried out to reveal the importance of prior knowledge and experience coupled with cognitive goal and its relevance to human eye gaze. E.g. drivers tend to look at the sides close to intersection (for sign boards) [6], batsmen focus at the bounce point of ball [33], people hold their gaze to the tip of the bottle/kettle while pouring the liquid [29] etc.

Gaze-directed displays are being used in attention-aware systems as Multi-layer displays [34], gaze-contingent display (e.g. the Tobii 1750 eye tracker for night vision surveillance, i.e.: www.tobii.se) and for intelligent tutoring [35] etc. In general, gaze-based systems are expected to reduce attentional workload of users, personalize the displayed information, attract user's attention—live or in teleconference, assist disabled and elderly persons etc.

3 Cyber Physical Healthcare Systems

3.1 Context-Awareness and Focus of Attention

Because of the huge number of actors involved in a healthcare system having complex communications, in addition to being context-aware, the system has to be capable of identifying the dynamic focus of attention and the exact actors/objects of relevance at any given instance. It is essential to note that attention-awareness encompasses much more than context-awareness. An attention-aware system is much more information-rich as compared to a simple context-aware one. Attention is typically defined as the set of processes enabling and guiding the selection of incoming perceptual information. This is specially required in healthcare systems, where the actors often come across multiple objects/actors at the same time, but not all of them are of interest at any given point in time. Things become even more difficult when the focus of attention switches among the given set of entities present in close proximity. Traditional sensing mechanisms like Radio-Frequency

Identification (RFID) sensors fail in such scenarios because of the various limitations: Ghost and/or Unread tags—in case multiple tags are read, the sensor may read a tag which is not actually present or conversely misses a tag originally present; Proximity—the sensors fail to work if placed closed to certain metal and/or liquid etc. Moreover, they are prone to attacks and require huge initial setup costs. In this chapter we exploit the fact that humans tend to "look" at the object of interest and hence we capture his eye gaze to identify his exact focus of attention.

3.2 Eye Gaze Tracking

Recent years has seen an increase in use of gaze-directed and first-person vision. Unlike traditional vision systems based on fixed cameras that observe people from the outside, first-person vision systems, with wearable gaze-capturing equipment, are able to work with data that relates directly to the user's interests and intentions.

The study of human working behavior reveals the importance of eye gaze in performing any task at hand. Human gaze and eye tracking offer insight into the cognitive processes involved in the tasks being carried out [7]. This was primarily revealed in the classical experiments by Yarbus [8], which forms the basis of further exploration [6]. Consider Fig. 3, which shows that eye gaze fixations and motions are strongly linked to the cognitive process which in turn is related to the goal/objective that the subject is trying to reach. Here the same person has been asked to look at the same painting and his eye motion trajectories are noted. These trajectories are different every time because the context in which the painting is viewed is different in each case.

Ever since human gaze has been shown to be directly linked to the human cognition and thought process [26, 36], it is increasingly being used in a number of domains centered around intelligent machines like design of intelligent homes [37], activity analysis [38], object and location recognition [39] to psychology and neuro-sciences [6] like study of human mind and identification of mental disorders etc. In each case the environment is captured from human point of view giving an insight into his exact focus of attention while performing any activity.

Due to the rapidly growing interest in gaze-based vision and eye gaze tracking, much effort is being applied to develop efficient, accurate, lightweight and easy to use eye gaze trackers. Moreover, as these systems are becoming more and more pervasive with a varying spectrum of users, including professionals, patients, children and even animals, more focus is being placed in making these systems more and more practical to handle. While initial systems were too intrusive and needed electrodes to be connected to the human body and/or wired contact lenses to be worn, current systems use imaging techniques to capture the eye gaze and comes with simple wearable cameras capable of capturing the scene from the wearer's point-of-view [26]. Some of the current and state-of-the-art wearable eye-gaze trackers are shown in Fig. 4 e.g. gaze-based camera used at "Quality of Life Technology" Centre (QoLT) [40], WearCam [41] and EyeSeeCam [30] etc.

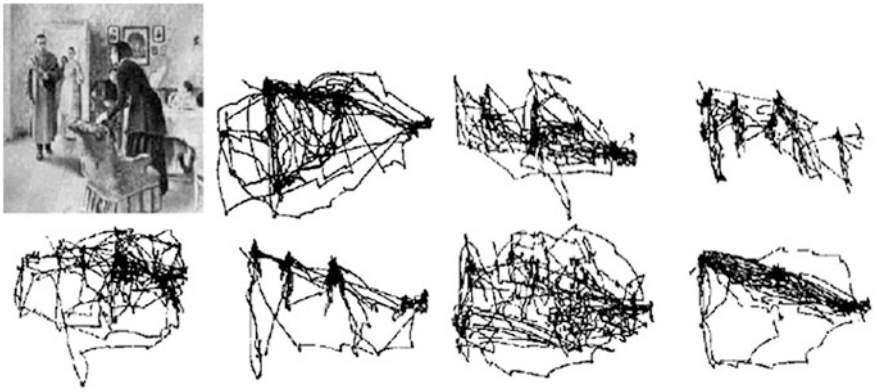


Fig. 3 Eye motion trajectories for a person looking at a painting in different context [8]

Fig. 4 Current eye-tracking systems

The EyeSeeCam, which is used in this work and shown in Fig. 4 (top row third) is described in Fig. 5 extends the technologies of eye trackers and head-mounted cameras by an eye-controlled camera that is always directed to the target of gaze. It is continuously oriented to the user's point of view by the eye movement signals of a mobile Video-Oculography (VOG) device and the scene is captured by the camera. An off-the-shelf notebook is able to record eye movement data and video streams from both eyes at up to 600 Hz as well as images from a wide angle scene camera and a gaze-driven camera with spatial resolutions of up to 752 × 480 pixels. It therefore captures a highly resolved image of what the user is looking at—just like the fovea.

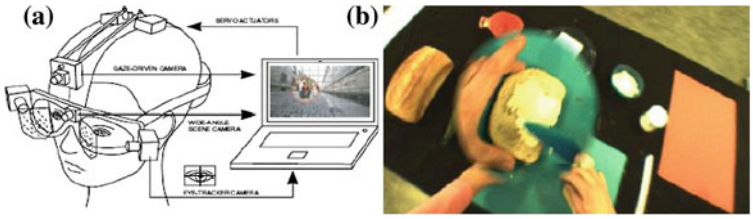


Fig. 5 EyeSeeCam—the eye gaze tracking device **a** Communication with computing device [30] **b** Camera output—the actor's perspective and focus of attention

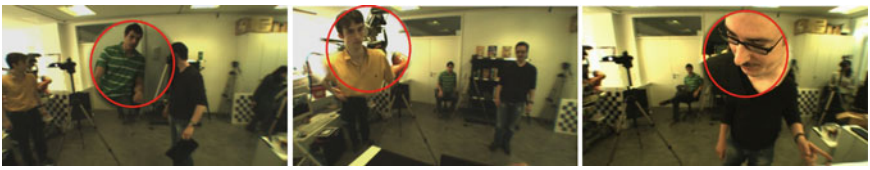


Fig. 6 Varying focus of attention at varying instances

Typically, the information derived from eye gaze data is through fixations, which indicate zero-motion of eyes (or pauses in eye movement), fixation behavior over time (e.g. spontaneous, task-relevant, orientation of thought looking, intentional manipulatory looking etc.) and variations in pupil size (indicating workload, complexity of task undertaken etc.). Consider Fig. 6 that shows the wearer's focus of attention switching among the group of personnel present in the environment. The gaze-based camera (EyeSeeCam in this case) is capable of capturing and highlighting the gaze direction at any given instance. The fixation duration may be noted to trigger the occurrence of desired event.

4 System Model and Methodology

The actor is equipped with a wearable eye tracker linked to a smart device. As soon as he enters the work environment, the gaze-directed camera starts getting the visual input and captures the scene where he is looking. The data is transmitted to his device which keeps checking for validity of data. Except in case of location identification, we use the eye-gaze fixation durations (i.e. pauses in eye scanning process) and consider a data as valid only if the fixation duration exceeds a threshold. For location identification a frame is selected periodically. Consider Fig. 7 that shows a high level model of the system, which links the life cycle, conceptual framework and communication stack of context-aware systems with the devices and logical sequencing at each and every step and phase.

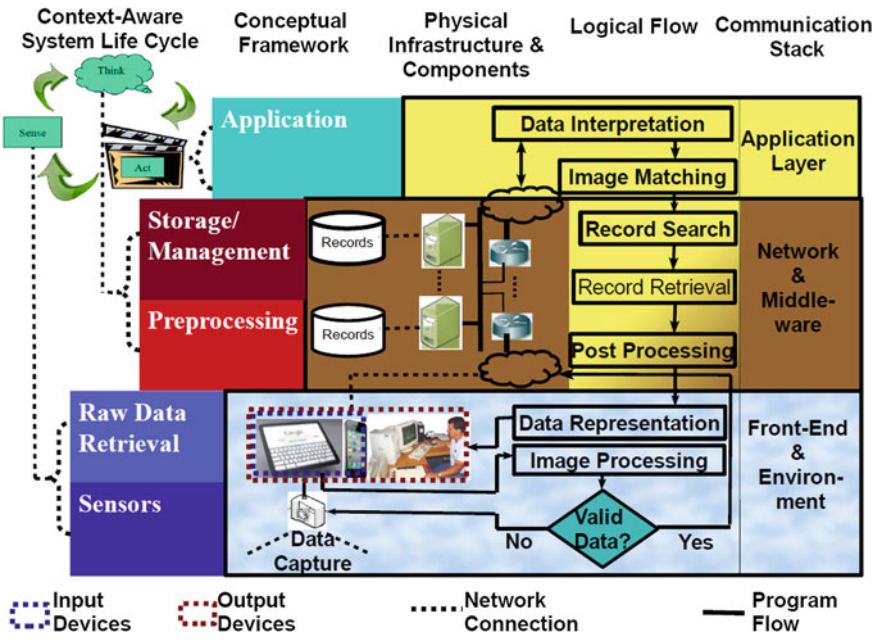


Fig. 7 System model

The modules are detailed below:

1. *Data Capture and Preprocessing:* The actor is equipped with a wearable eye tracker linked to a smart device. As soon as he enters the work environment, the gaze-directed camera starts getting the visual input and captures the scene where he is looking. Except in case of location identification, we use the eye-gaze fixation durations (i.e. pauses in eye scanning process) and consider a data as valid only if the fixation duration Φ exceeds a threshold θ. Hence we use the following equation to identify Validity υ of Object Ω being focused. For location identification, data captured every nth second or fth frame is used.

$$\upsilon(\Omega) = \begin{cases} 1 & \text{if } \Phi > \theta \\ 0 & \text{otherwise} \end{cases}$$

Consider Fig. 8, which shows the eye motion velocities in x, y and z direction as captured by EyeSeeCam and the fixations and saccades identified.

2. *Data Interpretation and Image Matching:* As soon as, a valid object is found, it is required to identify the scenario needed to be activated. Here, the personnel may be communicating with a person, requiring the face recognition and person identification module to be activated. He may be viewing the data on his smart device, for which the required "Tab" of information

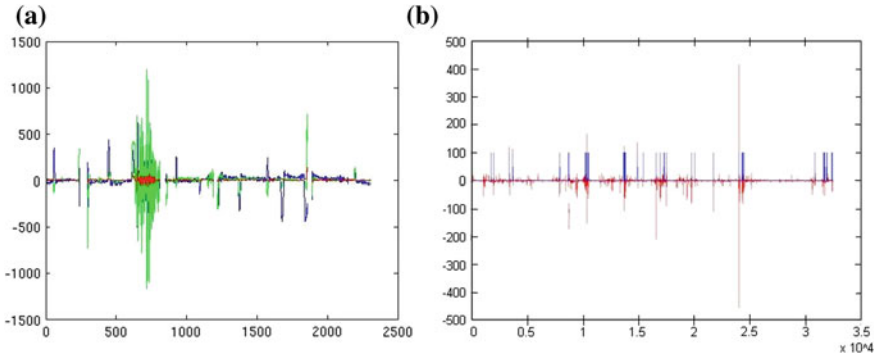


Fig. 8 **a** Eye motion trajectories **b** Fixations and saccades identified

needs to be loaded. Finally, in case of location identification, the environment model needs to be interacted with.

3. *Record Search and Retrieve:* This module is responsible for mapping the image information with the records physically stored on the servers. Database queries are used for extracting the desired record.
4. *Post-processing and Data Representation:* These modules convert the data into desired format according to the device used.

5 Experiments and Results

Although, the scheme presented is general and can be applied in any CPS like intelligent homes, life-critical systems e.g. air/road traffic control, Advance Driver Assistance System (ADAS), robotics etc., a setup of healthcare pervasive system is considered for experiments and the viability is shown through demonstration scenarios of person and menu identification and location recognition detailed below:

- Person Recognition: Attending a single person among a group of people is very common in a health care scenario e.g. a doctor attending patients in a ward or a nurse handing over medicines to clients in a multi-patient room. The conventional technique of having identification bands on patients and tags on devices won't work here, because multiple Radio-Frequency Identification (RFID) tags can be read simultaneously, and it is not possible to distinguish the relevant one from the rest. We capture the image of the person focused and use the Pittpatt Face Recognition SDK [42] to detect and identify the faces [43]. The algorithm detects the presence of faces in image(s), followed by retrieval of landmarks (like nose and eyes etc.) which are then compared to other faces in a gallery of images for recognition. For this

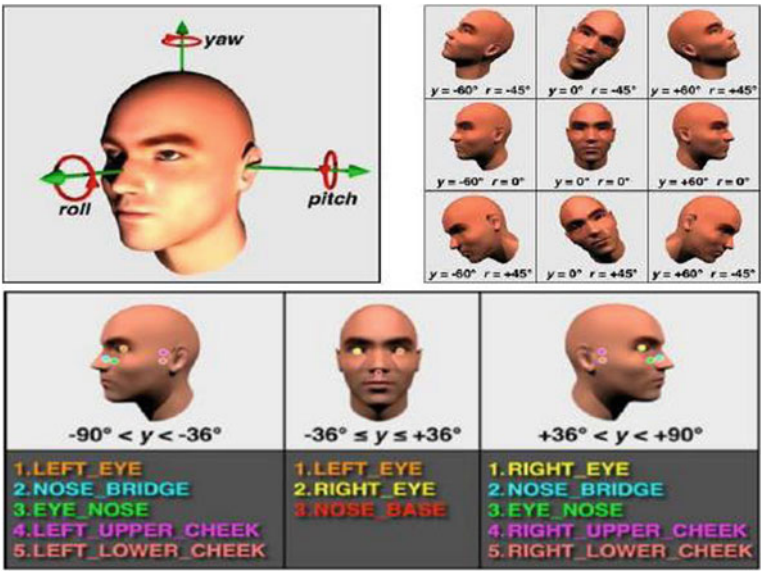


Fig. 9 Visual representation of face coordinate planes and the landmarks [42]

purpose a similarity matrix is computed, which is a way of storing the similarity scores between all subjects in a query gallery and all subjects in a target gallery. Subject-to-subject similarity scores and face-to-face similarity scores can both be obtained from a similarity matrix. Consider Fig. 9 (downloaded from [42]) that shows a visual representation of the yaw, pitch, and roll coordinate planes (Top) and the landmarks available in each yaw sector (Bottom).

- Location Identification: A number of location identification techniques exist like GPS satellites, mobile phone towers, badge proximity sensors, tags etc. However, these sources are neither 100 % accurate nor sufficient, especially when navigating closed and congested environments like hospital rooms. Moreover, they need expensive and huge setups and/or intrusive signals and radiations. Our technique using gaze-based vision and image matching is inexpensive, robust and practically unobtrusive as it comprises of image capturing equipment without contaminating the sensitive hospital environment with radiations and signals. This module utilizes a pre-stored neighborhood graph [44] of the environment, generated off-line. The input to our algorithm is a video or a set of images captured via an unconstrained camera traversal. The algorithm works purely on image measurables; hence no prior information of camera and/or scene parameters is desired. The algorithm uses the image set to form a 3D-tessellation around the object viewing space. This is achieved by developing a fuzzified neighborhood graph wherein the images represent the nodes. The strength of edges between the nodes depends

Sample image SIFT keypoints overlaid Color histogram

Fig. 10 A sample image from PASCAL VOC 2011 challenge [46] with it's SIFT points and histogram

on the proximity and appearance similarity of the constituent images. A detailed account of the model generation steps is given below.

We've used the image histogram and the Scale Invariant Feature Transform (SIFT) [45] as distinct features of our image. We begin with computing the features of all the reference images in the dataset. SIFT being scale and rotation invariant, and robust to changes in illumination and viewpoints provide a strong feature to represent the images in our scenario. It allows generation of a graph with precise information of the neighboring viewpoints, thus allowing us to remove any constraint on position and motion of camera and compensate for any motion of object during the data capture. The image histogram, on the other hand, provides us with the global image content and reinforces the proximity constraint. Consider Fig. 10 that shows a sample image with it's SIFT points and histogram.

Once the features are computed, the next step is to identify the corresponding features across the images. This allows us to compute the distance metrics which are then used to generate the final neighborhood graph.

Since, SIFT gives us a feature-by-feature local correspondence, hence we first compute a point-to-point match by calculating the dot products between the unit vectors and taking the inverse cosine of the resultant dot product. We then use these point correspondences to calculate the overall "distance" between the images. Consider Fig. 11 that shows the corresponding feature points for a pair of images.

These point correspondences are used for calculating image distances. Hence the set of feature points in the Image I_k is given by $F_k(x_i, y_i)$, with $k = 1, \ldots, n$, where n is the number of images in the dataset and $i = 1, \ldots, j$ where j is the number of key-points in kth image. Consequently, the distance d_f between two images I_m and I_n with c corresponding points is given as in Eq. (1). Here, I^x_{mj} and I^y_{mj} represent the x- and y-coordinates of the jth feature of mth image respectively where the points I_{mj} and I_{nj} correspond to each other.

$$d_f(\mathrm{I_m}, \mathrm{I}_{rmn}) = \frac{\sum_{j=1}^{c} \sqrt{\left(I^x_{mj} - I^x_{nj}\right)^2 + \left(I^y_{mj} - I^y_{nj}\right)^2}}{c}, \tag{1}$$

Fig. 11 *Left* Features identified and *Right* Correspondences marked for a pair of images

We use histogram intersection (d_h) [47] as a metric to compute the image similarity among images, which has been shown to be an effective representation for color-based models. Following the convention in [4], let H_k be the histogram of image I_k. Both histograms consist of b bins and the ith bin for i = 1, …, b is denoted by h^i_m and h^i_n. Then:

$$d_h(H_m, H_n) = \sum_{i=1}^{b} \min(h^i_m, h^i_n). \tag{2}$$

Figure 12 shows three sample images with their histograms and corresponding intersection among image pairs. Note that similar images (aka proximally close) give a higher intersection area (286,830 for images at viewpoint 1 and 2) as against different images (aka proximally distant) (240,705 for images at Viewpoints 2 and 3).

The features and distance metrics are combined to generate a fuzzified neighborhood graph that forms a view-centric representation of the object. Here the nodes denote the images while edges represent the neighbors, where neighborhood is defined in accordance to the viewpoint-proximity. Moreover, instead of having a binary true/false relation representing presence/absence of a link between images, we've used a fuzzy variable indicating "closeness" of views. Hence "close" viewpoints will have a stronger edge between them as compared to neighbors which are "far". Non-neighbors will have a link value (l) of zero.

For an image dataset with N images, an $N \times N$ sparse matrix is created. As compared to adjacency list or matrix, this allows ease of data handling in subsequent steps along with low space and time complexity, which is very critical for huge datasets with enormous, non-linked viewpoints.

The procedure for identification of neighbors is multi-fold. In the first pass, we find the distance between each image pair (I_m, I_n) by identifying the Euclidean Distance (d_f) and Histogram Intersection (d_h). Next, the two distance metrics are combined to form a cumulative distance metric as $d_{m,n} : \{d^f_{m,n}, d^h_{m,n}\}$. The distance between the image pair aka "closeness" gives us the strength of link ($l_{m,n}$) between them. Hence, images can be related to each other as {very close, close, somewhat close, not close} depending on the link strength. To define closeness of images, we

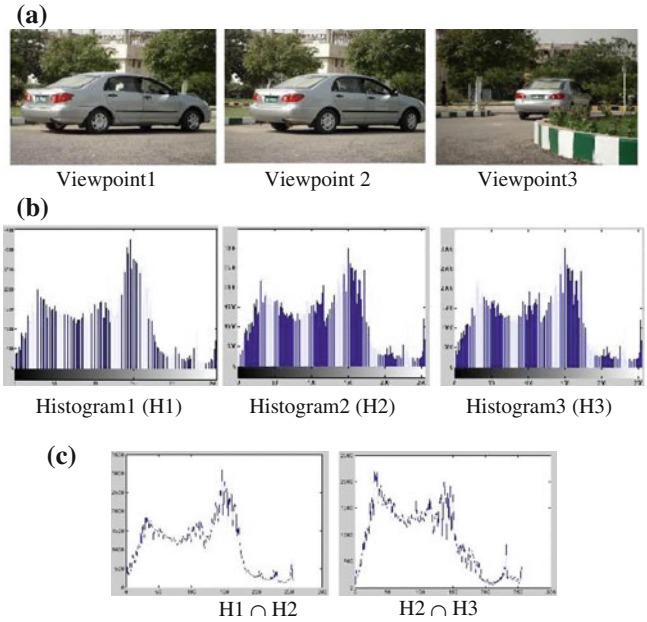


Fig. 12 **a** Images of a car from three different viewpoints. **b** Corresponding color histogram. **c** Histogram intersection for image pairs

begin with identifying, for each image, the pair having the minimum cumulative distance *d*. Considering this attribute as our seed point, we expand the region to include all those images in the neighborhood block, whose distance falls within a threshold relative to the minimum. This accounts for the out-of-plane images and handles arbitrary viewpoints. The "closeness" of neighbors may be defined using the membership functions low, medium and high. Hence, the link between image *m* and *n* is given as $l_{m,n}$ specifying the link strength depending on the proximity of viewpoints. Finally, video-continuity is used to form links between temporally adjacent images. Hence $l_{m,m+1} = 1 \; \forall \; m$. Consider Fig. 13 that shows the fuzzy sets defined for distance metrics and closeness.

Our approach for developing the topological structure of the images in the dataset provides ease of traversing while using video sequences. Let, at any time instant (t), a test frame (f_j) matches with the *i*th image in the neighborhood graph. Then, once a node has been detected at time t, for a true positive, there is an increased likelihood of observing the next test frame f_{j+1} at time $t + 1$ in the same proximity. We use the Bayesian image analysis to incorporate prior knowledge or beliefs into the recognition process. If this prior description of the neighborhood graph is given as N(x), the spatial distribution of this description, given the observed image f_j^t at time t, is:

$$p(N(x)|f_j^t) \propto l(f_j^t|N(x))p(N(x)), \tag{3}$$

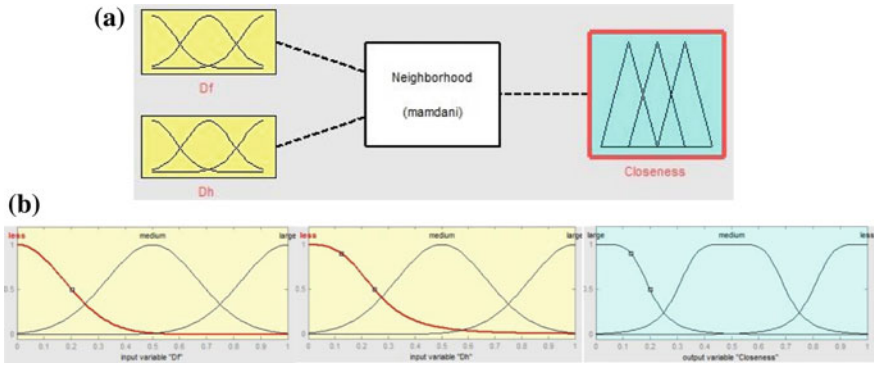


Fig. 13 Membership functions and model for the neighborhood graph. **a** The fuzzy model for neighborhood calculation. **b** Input variables: Euclidean distance (d_f) and histogram intersection (d_h) output variable: closeness

Fig. 14 Attention information for person identification, menu identification and location identification in a cyber healthcare system

where, $p(N(x)|f_j^t)$ is the posterior distribution and $l(f_j^t|N(x))$is the likelihood. Given one neighborhood graph for each object, only the correct object follows a smooth trajectory along the graph and the others suffer from discontinuities.

- Activation of Menu Tabs: Once a patient is identified, his information is loaded on the personnel's device in the form of separate tabs for personal details, radiology reports, appointments' schedule and diet plan etc. The practitioner may be interested in viewing only a particular subsection of patient's history—the "focused" one. We use template matching to match the pre-stored tabs with the current input image obtained by his gaze direction. For this purpose we use the Open Source function of OpenCV [48], which down samples both target and source images and calls *cvMatchTemplate()* on them. The specified number of best locations are found and for

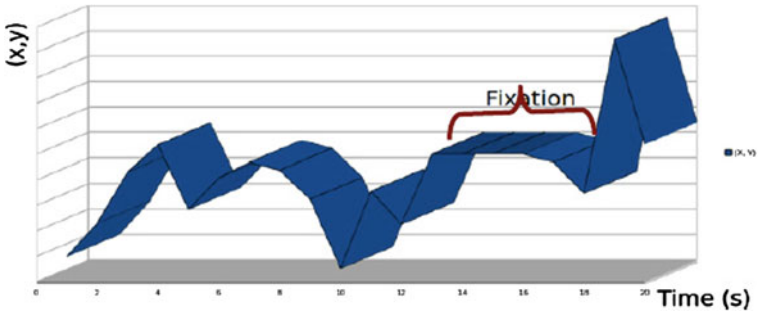


Fig. 15 Time-position graph representing eye-gaze trajectory for identifying focus of attention

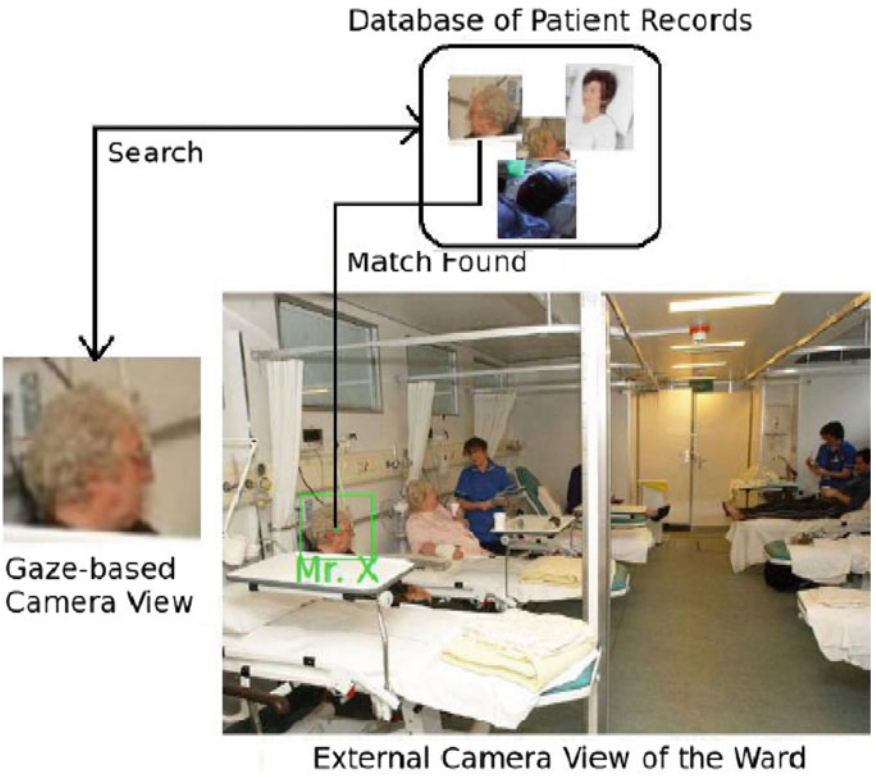


Fig. 16 A patient "Mr. X" identified just by having the practitioner look at him

each point where a maxima was located, original source image is searched in a defined window in both x and y direction. If match score is above the threshold then the location and score is saved.

The three scenarios are depicted in Fig. 14. The image matching results are marked as green boxes. We were able to generate the neighborhood graph and identify the locations with precisions of 97.86 and 80 % respectively. The faces were identified with an accuracy of about 95 % with the person recognition falling to 78.3 %. The template matching allowed menu identification with a precision of 98.53 %. In each case the focus of attention was correctly identified by fixation duration over 90 % of the time. The 10 % accounts for unintentional gazes at irrelevant objects.

Consider Fig. 15 that shows a sample eye gaze trajectory generated by a medical practitioner while navigating a ward. The fixation is identified and the image is used for matching and patient identification. Once the focus of attention is signaled, the image is extracted and used for person identification. Figure 16 shows the face recognition results for the patient-looked-at. The green box identifies a match of the person from the database in the given scene. His ID is mentioned alongside.

6 Conclusion

In this chapter, a technique for identifying the focus of attention in a Cyber Physical System is presented via a context-aware healthcare system. We propose to use first-person vision, obtained through wearable gaze-directed camera that can capture the scene through the wearer's point-of-view. We use the fact that human cognition is linked to his gaze and typically the object/person of interest holds our gaze. Upon identification of the point of interest of the person wearing the gaze-based camera, the image is processed to search for the face of the person or a medical tab on the handheld device, which is then used to recognize the identity of the patient/tab looked at. The record is then retrieved and displayed on the practitioner's hand-held device. Moreover, the images of his surroundings are periodically captured and compared with the environment model to identify his exact location. We argue that our technique is robust and works well in the presence of noise and other distracting signals, where the conventional techniques of IR sensors and tagging fail. Moreover, the technique is unobtrusive and does not pollute the sensitive hospital environment with unnecessary signals. Our approach is general in that it may be applied to CPS other than healthcare and can be shown to work equally well in other office and industrial scenarios and also in intelligent homes.

References

1. B.H. Krogh, *Cyber Physical Systems: The Need for New Models and Design Paradigms*. Presentation Report, Carnegie Mellon University
2. R. Rajkumar, *CPS Briefing* (Carnegie Mellon University, Pittsburgh, 2007)

3. J. Shi, J. Wan, H. Yan, H. Suo, A survey of cyber-physical systems, in *International Conference on Wireless Communications and Signal Processing*, Nanjing, China, 9–11 Nov 2011
4. B.P. Bailey, J.A. Konstan, On the need for attention-aware systems: measuring effects of interruption on task performance, error rate, and affective state. Comput. Hum. Behav. **22**, 685–708 (2006)
5. G. Schirner, D. Erdogmus, K. Chowdhury, T. Padir, The future of human-in-the-loop cyber-physical systems. Computer. IEEE Computer Society Digital Library (2012)
6. M. Hayhoe, D. Ballard, Eye movements in natural behavior. Trends Cogn. Sci. **9**(4), 188–194 (2005)
7. M.F. Land, Vision, eye movements, and natural behavior. Vis. Neurosci. **26**, 51–62 (2009)
8. A.L. Yarbus, *Eye Movements and Vision* (Plenum, New York, 1967)
9. W. Wolf, *Cyber-Physical Systems* (IEEE Computer Society, Los Alamitos, 2009)
10. E.A. Lee, Cyber-physical systems—are computing foundations adequate*? NSF Workshop on Cyber-Physical Systems: Research Motivation, Techniques and Roadmap*, Austin, TX, 16–17 Oct 2006
11. L. Sha, S. Gopalakrishnan, X. Liu, Q. Wang, Cyber-physical systems: a new frontier, in *IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing* (2008)
12. P. Tabuada, Cyber-physical systems: position paper, in *Symposium Conducted at the NSF Workshop on Cyber-Physical Systems*, Austin, TX, October 2006
13. Y. Tan, S. Goddard, L.C. Pérez, A prototype architecture for cyber-physical systems, in *SIGBED Review, Special Issue on the RTSS Forum on Deeply Embedded Real-Time Computing,* vol. 5(1) (2008)
14. L.T. Kohn, J. Corrigan, *To Err Is Human: Building a Safer Health System* (National Academy Press, Washington, 2000)
15. J. Bohn, F.C. Gärtner, H. Vogt, Dependability issues of pervasive computing in a healthcare environment, in *SPC 2003* (2003), pp. 53–70
16. J.E. Bardram, H.B. Christensen, A.K. Olsen, *Activity-Driven Computing Infrastructure–Pervasive Computing in Healthcare*. Department of Computer Science, Center for Pervasive Computing, University of Aarhus 2004, CfPC-2004-PB-65 (2004)
17. D. Zhang, Z. Yu, C.Y. Chin, Context-aware infrastructure for personalized healthcare. Stud. Health Technol. Inform. **117**, 154–163 (2005)
18. J.H. Jahnke, Y. Bychkov, D. Dahlem, L. Kawasme, Context-aware information services for health care, in *Proceedings of International Workshop on Modelling and Retrieval of Context*, vol. 114 (2004)
19. M. Baldauf, S. Dustdar, F. Rosenberg, A survey on context-aware systems. Int. J. Ad Hoc Ubiquitous Comput. **2**(4), 263–277 (2007)
20. C. Orwat, A. Graefe, T. Faulwasser, Towards pervasive computing in health care—a literature review. BMC Med. Inform. Decis. Mak. **8**, 26 (2008)
21. N. Bricon-Souf, C.R. Newman, Context awareness in health care: a review. Int. J. Med. Inform. **76**(1), 2–12 (2007)
22. M. Miraoui, C. Tadj, C. Ben Amar, Architectural survey of context-aware systems in pervasive computing environment. Ubiquitous Comput. Commun. **3**(3) 68–76 (2008)
23. D.O. Kang, H.J. Lee, E.J. Ko, K. Kang, J. Lee, A wearable context aware system for ubiquitous healthcare, in *Proceedings of the 28th Annual International Conference of the IEEE on Engineering in Medicine and Biology Society*, New York, NY (2006)
24. C. Roda, J. Thomas, Attention aware systems, in *Encyclopaedia of HCI*, ed by C. Ghaoui (IDEA Group, Hershey, 2005)
25. C. Roda, J. Thomas, *Attention Aware Systems: Theory, Application, and Research Agenda* (Computers in Human Behavior, Elsevier, 2006)
26. A. Toet, Gaze directed displays as an enabling technology for attention aware systems. Comput. Hum. Behav. **22**(4), 615–647 (2006)

27. G. Boening, K. Bartl, T. Dera, S. Bardins, E. Schneider, T. Brandt, Mobile eye tracking as a basis for real-time control of a gaze driven head-mounted video camera, in *Symposium on Eye Tracking Research and Applications* (2006)
28. T. Brandt, S. Glasauer, E. Schneider, A 3rd eye for the surgeon. J. Neurol. Neurosurg. Psychiatry **77**(2), 278 (2006)
29. M. Land, N. Mennie, J. Rusted, The roles of vision and eye movements in the control of activities of daily living. Perception **28**(11), 1311–1328 (1999)
30. E. Schneider, T. Villgrattner, J. Vockeroth, K. Bartl, S. Kohlbecher, S. Bardins, H. Ulbrich, T. Brandt, EyeSeeCam: an eye movement-driven head camera for the examination of natural visual exploration. Ann. N. Y. Acad. Sci. **1164**, 461–467 (2009)
31. P. Wagner, K. Bartl, W. Günthner, E. Schneider, T. Brandt, H. Ulbrich, A pivotable head mounted camera system that is aligned by three-dimensional eye movements, in *Symposium on Eye tracking research and applications*, New York, NY, USA (2006), pp. 117–124. doi:10.1145/1117309.1117354
32. M.F. Land, M. Hayhoe, In what ways do eye movements contribute to everyday activities? Vision. Res. **41**, 3559–3565 (2001)
33. M.F. Land, P. McLeod, From eye movements to actions: how batsmen hit the ball. Nat. Neuroscience. **3**, 1340–1345 (2000)
34. Deep Video Imaging Ltd. (2004), Interactive dual plane imagery, http://www.deepvideo.com
35. C.R. Beal, Adaptive user displays for intelligent tutoring software. Cyber-Psychol. Behav. **7**(6), 689–693 (2004)
36. G.J. Zelinsky, R.P.N. Rao, M. Hayhoe, D.H. Ballard, Eye movements reveal the spatiotemporal dynamics of visual search. Psychol. Sci. **8**(6), 448–453 (1997)
37. Sensing at QoLT, http://www.cmu.edu/qolt/Research/projects/current-projects/sensing.html#sensing%20project
38. L. Sun, U. Klank, M. Beetz, EYEWATCHME—*3D* hand and object tracking for inside out activity analysis, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 20–25 June 2009
39. H. Kang, A.A. Efros, M. Hebert, T. Kanade, Image matching in large scale indoor environment, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop on Egocentric Vision,* Miami, FL, USA, 20–25 June 2009
40. T. Kanade, First-person, inside-out vision. Keynote speech, in *Proceedings of 1st Workshop on Egocentric Vision, in Conjunction with CVPR,* Miami, FL, USA, 20–25 June 2009
41. B. Noris, J.B. Keller, A. Billard, A wearable gaze tracking system for children in unconstrained environments. Comput. Vis. Image Underst. **115** (4) 476–486 (2011)
42. Face Recognition SDK at Pittsburgh Pattern Recognition Group (2011), http://www.pittpatt.com.. Accessed 23 August 2011
43. S. Noor, H.N. Minhas, Aqeel-ur-Rehman, Using gaze-directed vision to identify focus of attention in pervasive healthcare systems, in *IEEE 14th International Multitopic Conference INMIC*, SSUET, Karachi, Pakistan (2011)
44. H.N. Minhas, S H. Mirza, Y. Sheikh, A. Jain, M. Shah, Model generation for video-based object recognition, in *Proceedings of 14th ACM Multimedia Conference at Santa Barbara*, USA (2006)
45. D.G. Lowe, Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60**(2), 91–110 (2004)
46. PASCAL (2011), Pattern analysis, statistical modeling and computational learning visual object classes challenge—VOC2011, http://pascallin.ecs.soton.ac.uk/
47. M.J. Swain, D.H. Ballard, Color indexing. Int. J. Comput. Vision **7**(1), 11–32 (1991)
48. Fast Match Template (2010), Function of OpenCV at OpenCVWiki, http://opencv.willowgarage.com/wiki/FastMatchTemplate. Accessed 19 April 2010

A Comparison of Pulse Compression Techniques for Ranging Applications

Aamir Hussain, Zeashan H. Khan, Azfar Khalid and Muhammad Iqbal

Abstract In this chapter, a comparison of Golay code based pulse compression (GCPC) technique with the Neuro-Fuzzy based pulse compression (NFPC) technique is demonstrated for ranging systems. Both of these techniques are used for the suppression of range side lobes that appear during the pulse compression process of the received echo pulse at the receiver for target(s) detection. Golay code is a pair of complementary codes and has an inherent property of zero side lobes when the two auto-correlation results of the complementary code pair are added. On the other side, neural network based pulse compression techniques are also developed to reduce the range side lobes. Both the techniques are different in nature but they share the common objective of range side lobe suppression in target detection. The differentiation parameters chosen for the comparison of GCPC and NFPC techniques include the computational complexity, range side lobe suppression levels, noise rejection capability, Doppler tolerance capability, range resolution capability as well as the training and convergence requirements of these pulse compression techniques. All these comparison criteria are found to determine the overall performance measures of the pulse compression techniques for ranging applications especially in case of detection and ranging of multiple closely spaced and weak targets. This comparison may be useful for a system designer to select a particular type of pulse compression technique for a specific ranging application.

Keywords Phase coded range estimation/detection · Pulse compression · Range side lobe suppression · Neuro-fuzzy network · Golay codes · Range resolution

A. Hussain (✉) · Z. H. Khan
National University of Sciences and Technology (NUST), Islamabad, Pakistan
e-mail: aamirhussain@ee.ceme.edu.pk

A. Khalid
Muhammad Ali Jinnah University, Islamabad, Pakistan

M. Iqbal
COMSATS Institute of Information Technology, Wah Cantt, Pakistan

Z. H. Khan et al. (eds.), *Computational Intelligence for Decision Support in Cyber-Physical Systems*, Studies in Computational Intelligence 540,
DOI: 10.1007/978-981-4585-36-1_5,

1 Introduction

Ranging sensors and systems are being used in multiple roles in biomedical, industrial, aerospace and military cyber physical systems. Ranging systems are used both for short range and long range applications; e.g., radio-altimeters on aerial systems and range finder in robotic vehicles, echogram in medical diagnostics applications at the close-range end (where the range may be a few feet or even less) and early warning and air traffic control radars for longer range to name few. In modern ranging systems, phase coding of the transmitted signal is carried out to get better range resolution capability [1–5]. Range resolution is typically defined as the capability of a ranging system to resolve two closely separated objects in range [6]. It has special and critical importance in aerial radar systems, medical diagnostics, air traffic control, weather radar, early warning radars and in robotics where the radars mounted on robots have to make very important decisions on the basis of ranges of the closely spaced objects and obstacles etc. Such systems communicating while tracking an object or mapping an area will form a robotic cyber physical system (CPS) with different range sensors mounted on various mobile vehicles e.g. for map building based on multi sensor data fusion (MSDF).

For a phase coded pulsed ranging system (PRS), range resolution is directly proportional to the chip width (time taken by 1 bit) of the coding sequence [1–3]. Barker codes, PN (pseudo noise), Kasami and Gold sequences are the conventional phase coding signals used in the ranging applications to get enhanced range resolution capability [1–5]. A train of phase coded pulses is transmitted in the direction where one or more targets are expected to be present. Target detection is carried out by the correlation of the received echo signal with the phase coding sequence at baseband [6–8], and the process is called pulse compression. The conventional phase coding signals in ranging applications have a problem that side lobes appear in their pulse compression at the receiver at baseband [9, 10]. These side lobes may result in false alarms, and weak targets may get masked or suppressed in the side lobes [11]. Range resolution capability of the ranging system is also impeded which causes hindrance in the detection and ranging of multiple closely spaced targets [5].

One of the main purposes of waveform design for pulse compression in ranging systems is to solve the dilemma between the range resolution and the pulse length [12]. Pulse compression processing is the most significant factor in determining the performance of high-resolution and highly sensitive radars. For instance, synthetic aperture radar (SAR) always contains a high range resolution pulse compression in addition to downward-looking rain measuring radar with a range side lobe level of 55 dB [12]. In satellite-borne rain radar, very stringent requirement of 60 dB is posed on range side-lobe level; while the air traffic control (ATC) system requires the side-lobe lower than 55 dB under the main lobe level [12].

Significant efforts have been made to reduce, suppress or equalize range side lobes in phase coded ranging systems. Ackroyd et al. [13] have presented an optimal inverse filtering technique for minimizing the integrated side-lobe level (ISL)

in least square sense. Daniels [14] has proposed code inverse filtering for complete side-lobe removal. But the performance of matched filter as well as the inverse filter degrades in the presence of a Doppler shifted return of the signal [15]. Bucci [16] has proposed a Doppler tolerant range (DTR) side-lobe suppression scheme. Mudukutore et al. [15] use Pulse compression to improve range resolution while maintaining a high duty cycle. Gartz [17] has generated uniform amplitude complex code sets with low correlation side lobes. Suto et al. [18] have introduced a time side lobe reduction technique for binary phase code pulse compression. Wu et al. [19] have introduced a chaotic phase code for Radar Pulse compression. Kai et al. [20] discuss a new wave form with high range resolution. Chandrasekhar et al. [21] present waveform coding for dual polarization weather radars. Kusk et al. [22] presents the azimuth phase coding for range ambiguity suppression. Shen et al. [23] propose window functions for reducing the side-lobe levels and obtaining high range resolution. Jenshak et al. [24] seek to minimize the estimation error of a range profile directly. Anelson [25, 26] uses ultra-wideband noise waveforms for side lobe suppression, and applies mismatched filtering to suppress the side-lobes of random noise radar.

Almost, all the techniques developed for range side lobe suppression have some limitations in them. For example, in most of the techniques, tapering of the matched filter response is done by weighting the transmitted waveform, the matched filter response or both in frequency or amplitude. For these tasks additional signal processing resources are required [14, 16]. Main lobe peak or processing gain of the output is reduced by these methods. Also, main lobe width of the matched filter output is widened which results in reduced range resolution [16]. In the techniques involving some additional filtering, the filter coefficients vary from code to code and change as a function of code and filter length [14, 16, 24–26]. The filters and weighting functions reduce the signal to noise ratio (SNR) at the output. Also range side lobes are not completely suppressed in most of the cases [15, 17–23]. Most of the techniques are not optimal and result in performance loss [14–24]. Due to the limitations of the existing techniques, and given the importance of the range side lobe suppression problem, “range side lobe suppression” is an active area of research in ranging applications particularly involving the detection and ranging of multiple and weak targets.

Some researchers have devoted themselves to developing the pulse compression algorithms for this century’s advanced weather radar to meet the higher time and space resolution requirements [27–31]. Eventually, the main purpose of the pulse compression is to raise the signal-to-maximum side-lobe (signal-to-side-lobe) ratio (SSR) and decrease the integrated side-lobe level (ISL) which is defined [30] as to improve the detection and range resolution abilities of the radar system. Also, for a good pulse compression algorithm, certain performance must be considered, including the noise rejection and the Doppler tolerance performance [12]. Here, the side-lobes are unwanted by-products of the pulse compression process; for the correlation of a code, the side lobes are the portions of the output waveform non-matching with the code other than the output of matching the code. And the side-lobe level is the magnitude of the side lobe [13].

In this chapter we have made a comparison of the Golay code based and the Neural network based pulse compression techniques. The purpose of both the pulse compression techniques is to increase the main lobe signal level of the compressed pulse signal along with the suppression of side lobes [12, 31–33]. Golay code based pulse compression exploits the range side lobe suppression property of complementary Golay code pair to achieve zero level side lobes [34–41]. No additional processing of the output result of the compressed pulse is required to suppress or reduce the side lobes [34–41]. The computational complexity of the Golay code based pulse compression algorithm is a little more but it can be handled by using modern fast digital hardware [34–41].

Recently, on the other side, neural networks applied to pulse compression were proposed with their learning capabilities [12, 31–33]. Kwan and Lee [32] have employed a back propagation (BP) algorithm to realize pulse compression with a phase-coded waveform, and obtained a good result. Moreover, some other traditional algorithms such as direct autocorrelation filter (ACF), least squares (LS) inverse filter, and linear programming (LP) filter based on B13 code are developed for pulse compression in ranging systems. The traditional algorithms cannot achieve the requirements of high signal-to-side lobe ratio and low integrated side lobe level (ISL) [12, 31–33]. Also, the convergence speed of the BP and these traditional algorithms is inherently low [31–33] and sensitive to the Doppler frequency shift. To cope with the drawbacks associated with the neural network based pulse compression techniques, a novel solution to the problem of pulse compression was proposed which is a self-constructing neural fuzzy inference network (SONFIN) [12]. SONFIN is used to perform a B13 code with the sequence $\{1, 1, 1, 1, 1, -1, -1, 1, 1, -1, 1, -1, 1\}$ and a 20-element combined Barker code (CBC) expanded by combining known Barker code with the sequence [18]. The algorithm combines the Barker code with SONFIN to constitute the Neuro-Fuzzy pulse *compression (NFPC)* [13].

Simulations have demonstrated that the side lobe at the output of NFPC can be significantly decreased though not completed suppressed. The success of SONFIN is due to the combinations of the self-constructing neural fuzzy inference network and both the short, simple, ease-implementing B13 code and 20-element CBC, respectively. Moreover, while compared with traditional algorithms e.g. ACF, LS, LP, and BP, NFPC has shown better performance in terms of noise rejection ability, higher range resolution and superior Doppler tolerance. Another important advantage of NFPC is that it has higher convergence speed than BP algorithm [12]. These examining results lead NFPC to be very suitable for the high-resolution radar systems. But NFPC also has the disadvantage of computational complexity.

The comparison in this chapter is done in order to access the performances of the two pulse compression techniques. This comparison is based on the suppressed side lobe levels (SLS), computational complexity, noise rejection capability, range resolution capability, Doppler tolerance capability, maximum ranging capability and training and convergence requirements of these algorithms. This comparison of the two pulse compression techniques is useful for the system designer to select

a particular pulse compression technique as per application requirement. These criteria are found to play an active role in the target detection process [8–11].

The rest of this chapter is organized as follows. Section 2 describes the ranging problem briefly. In Sect. 3, we shall discuss briefly the Golay code based pulse compression for ranging applications. In Sect. 4, there is a very brief overview of neural networked pulse compression techniques for radar applications. In Sect. 5, we present the comparison of the Golay code based pulse compression with the neural networked pulse compression techniques for ranging applications. The chapter ends with conclusions in Sect. 7.

2 Detection and Ranging of Phase Coded Pulsed Signals

Generalized Likelihood Ratio Test (GLRT) detector is an appropriate option in case signal has unknown parameters. When employed as a detector/estimator for the detection of an echo of the transmitted signal s[n] embedded in white Gaussian noise, the GLRT can be represented as in Eq. (1) [3, 4, 8, 9, 11, 12].

$$\sum_{n=\hat{n}_o}^{\hat{n}_o+M-1} x[n]s[n-\hat{n}_o] \geq \gamma' \tag{1}$$

Expression (1) shows the correlation of $x[n]$ (the received noisy echo) with the delayed transmitted signal $s[n-\hat{n}_0]$ of length M. The maximum value of this correlation is obtained at $n=\hat{n}_0$, which is compared with a threshold γ'. If the threshold is exceeded, a signal (target) is declared to be present and its delay (range) is estimated as $\hat{n}_0$; otherwise absence of any target is declared [5–7]. In case multiple targets are present, multiple correlation peaks will exceed the threshold at different delays (ranges) [7].

3 Golay Code Based Pulse Compression (GCPC)

This technique exploits the side lobe suppression property of Golay code to achieve range side lobe suppression and thus enabling improved range resolution possible in ranging applications. Golay code is a pair of complementary codes, say code c and code k, each having length of 2^L where L is a positive integer [34, 35]. Golay codes have a useful property that addition of auto-correlation results of the two complementary Golay codes gives zero side-lobes [36–38]. The side lobes in the two correlation results being complementary to each others are canceled and completely suppressed as shown in Fig. 1 for a complementary code pair of length 64 [39–41]. In the ranging applications, the shift in the index of correlation peak gives the delay corresponding to the range of the object/target [41].

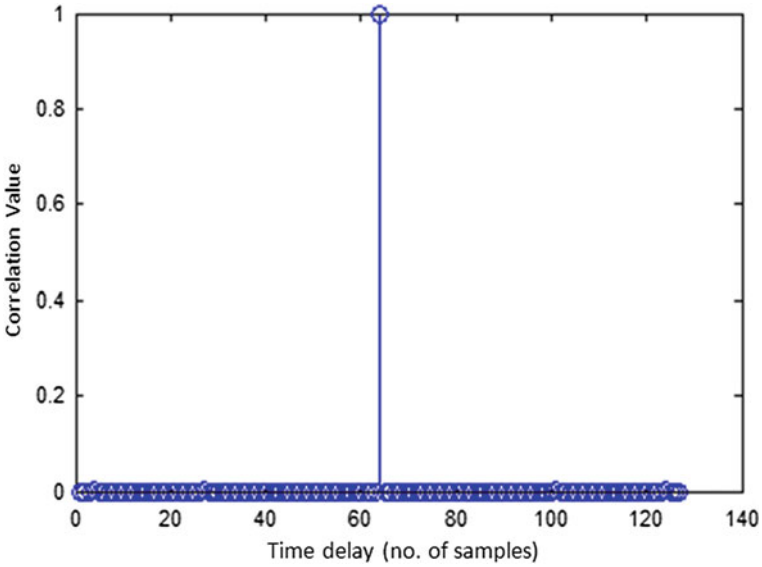


Fig. 1 Suppressed side lobes in the pulse compression of Golay code pair

The side lobe suppression property of Golay codes is well-known in literature [34–41]. Due to the absence of side lobes, false alarms are avoided and the weak targets may not get masked in the side lobes [1–11]. Improved range resolution is also achieved [1–11]. Golay codes have conventionally been used for channel coding in digital communications [34–36]. Using Golay codes in ranging applications requires some special considerations as we have to use a pair of complementary codes for phase coding the carrier signal.

Special considerations are required both at the transmitter side and at the receiver side to transmit, receive and then process the received Golay coded signal to achieve zero side lobes in the compressed pulse result. Customized signal processing algorithms are developed for phase coded pulse Doppler radar using Golay code as the phase coding signal. With the development of these algorithms, along with Doppler localization of moving targets, the range side lobes problem of the conventional radar has been overcome and improved range resolution is achieved.

3.1 Target Detection and Range Estimation: GCPC Example

Consider a single stationary target is present at a range of 6 km from the transmitter. Pulse coded with code-c is first transmitted towards the target, and then after some delay the pulse coded with code-k is transmitted. Length of each member (code c or code k) of the complementary Golay code pair is $M = 64$

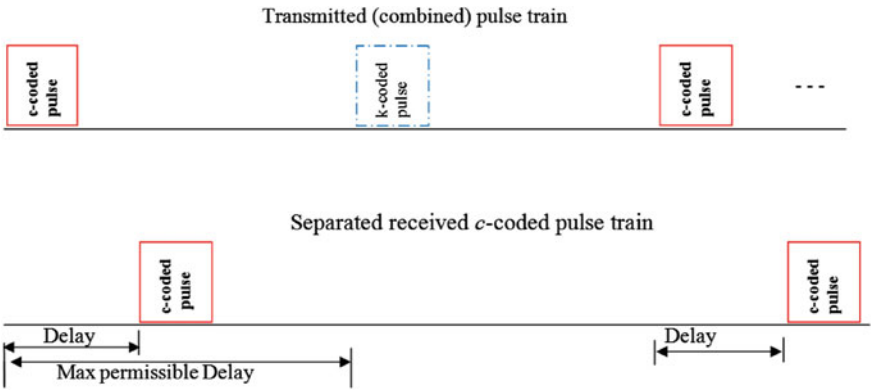


Fig. 2 Golay code based transmitted and received pulse trains

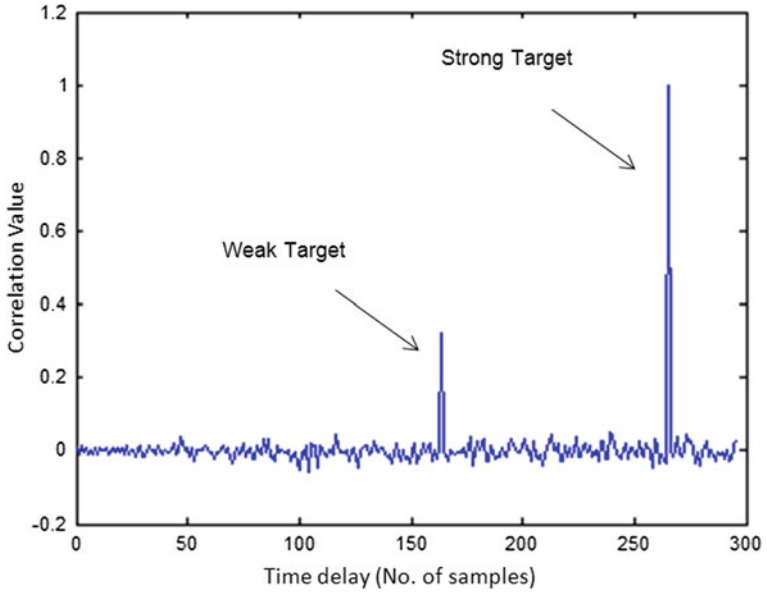


Fig. 3 Detection of a strong target at a range of 6 km, and a nearby weak target in the noisy environment

chips. Chip time is $T_c = 0.2\,\mu s$. Pulse width is $T' = MT_c = 12.8\,\mu s$. Time delay for 6 km range is $T = 2R/C = 40\,\mu s$. This time delay in terms of number of samples is $T_s = T/T_c = 200$ samples delay.

Correlation of the received c-coded pulse is carried out with code-c and the peak value of correlation is achieved at a delay $\hat{n}_0$ [1–11]. In this result, range side

lobes are present as this process involves auto-correlation of code-c as shown in Eq. (2).

$$r_1[n] = \sum_{\hat{n}_0}^{\hat{n}_0+M-1} c[n-n_0]c[n-\hat{n}_0] \tag{2}$$

The index of correlation peak ($200 + 64 = 264$ samples) gives the delay corresponding to the range of target; but this correlation has side lobes present. Similarly, the correlation of the received k-coded pulse is carried out with code-k and the transmission lag (delay between the transmissions of the two pulses) is compensated as shown in Fig. 2. This result also has range side lobes present as it involves the auto-correlation of code k as shown in Eq. (3).

$$r_2[n] = \sum_{\hat{n}_0}^{\hat{n}_0+M-1} k[n-n_0]k[n-\hat{n}_0] \tag{3}$$

Finally, the two correlation results are added to achieve the suppressed correlation side lobes. The final result of the pulse compression algorithm, Eq. (4), is free from range side lobes.

$$y(n) = \left[\sum_{n_0}^{n_0+M-1} c[n-n_0]c[n-\hat{n}_0] + \sum_{n_0}^{n_0+M-1} k[n-n_0]k[n-\hat{n}_0]\right] \tag{4}$$

The peak of this final correlation result is the true MLE (maximum likelihood estimate) of the delay (n_0) corresponding to the range to the target. Detection and ranging of a strong target along with a nearby weak target in the presence of noise using Golay complementary code is shown in Fig. 3. The signal strength and hence the correlation peak reduces proportionally after reflecting from a weak target.

4 Neural Network Based Pulse Compression

As discussed in the introduction section, neural network learning applied to pulse compression were proposed in [12, 31–33]. In order to achieve the requirements of high signal-to-side lobe ratio and low integrated side lobe level (ISL), the normal neural networks such as the back propagation (BP) networks are not recommended as they are sensitive to the Doppler frequency shift [31–33]. The use of SONFIN in the proposed NFPC scheme is obviously not the only choice. Other types of neural networks or fuzzy systems are also possible. However, our choice of SONFIN is based on several reasons. First, the SONFIN is a hybrid system of neural networks and fuzzy logic. With a fuzzy-inference-type structured network, the SONFIN can always achieve higher learning accuracy and convergence speed than normal neural networks [12]. Also, the IF–THEN-typed expert knowledge can be put into

or extracted from the SONFIN easily [12]. Second, as compared to the existing fuzzy neural networks, the SONFIN can perform both the structure and parameter learning simultaneously such that it can online construct itself on the fly dynamically [12]. As a result, it can always find itself a very economic size of network for a given learning task while comparing the other Neuro-Fuzzy [12]. Here, we briefly describe this technique for radar pulse compression.

4.1 Use of SONFIN to Process the Pulse Compression in a Radar System

The block diagram of the digital pulse compression system using NFPC is shown in Fig. 4 [12]. The Barker code generator generates the B13 code sequences or the 20-element CBC sequences, which are sent to RF modulator and transmitter. Received IF signals are passed through a band pass filter matched to the sub-pulse width and are demodulated by two detections, I_det and Q_det, with a local-oscillator (LO) signal at the same IF frequency, and then the in-phase (I) and quadrature (Q) channel echo signals are detected, respectively [12]. These echo signals are converted to digital form by analog-to-digital (A/D) converters under the system timing control that also clocks the Barker code to be transmitted. The digital form of the echo signals consists of the Barker code and interfering noise.

The NFPCs, which are implemented by the trained SONFIN, carry out the pulse compression based on the received sequence. Once, the echo sequence is matched with the transmitted Barker code, the output of each SONFIN will be +1 with one sub-pulse duration. When the SSR of the NFPC output is very high, the false alarm of the detector is reduced, and eventually the detection ability of the radar system is enhanced [12].

4.2 Structure of SONFIN

The structure of SONFIN for B13 code is shown in Fig. 5 below [12]. There are no rules initially in the SONFIN. They are created and adapted as online learning proceeds via simultaneous structure and parameter learning, so the SONFIN can be used for normal operation at any time as learning proceeds without any assignment of fuzzy rules in advance. This six-layered network realizes a fuzzy model of the following form:

Rule i: IF x_1 is A_1^i and…. and

$$x_n \text{ is } A_1^n \text{ THEN } y \text{ is } m_o^i + b_j^i x_j + \ldots\ldots$$

where, A_i^j is the fuzzy set of the i^{th} linguistic term of input variable x_j, m_0^i is the center of a symmetric membership function on y, and b_j^i is the consequent

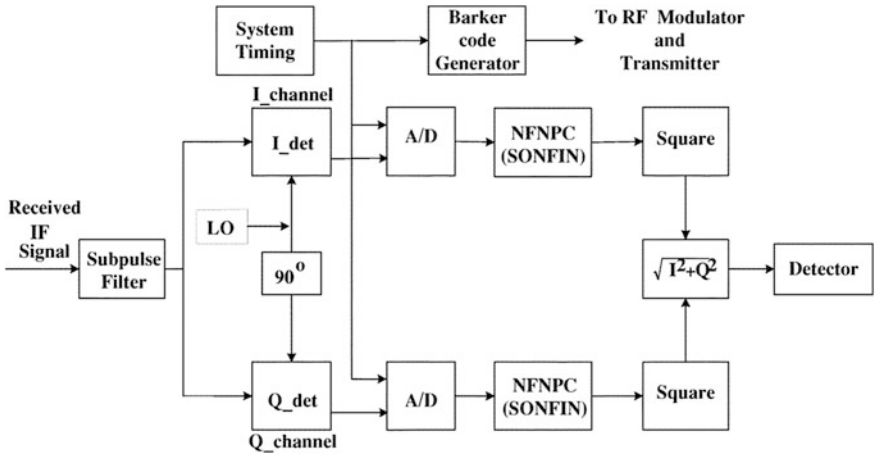


Fig. 4 Block diagram of the digital pulse compression system using Neuro-Fuzzy Networked Pulse Compression NFPC [12]

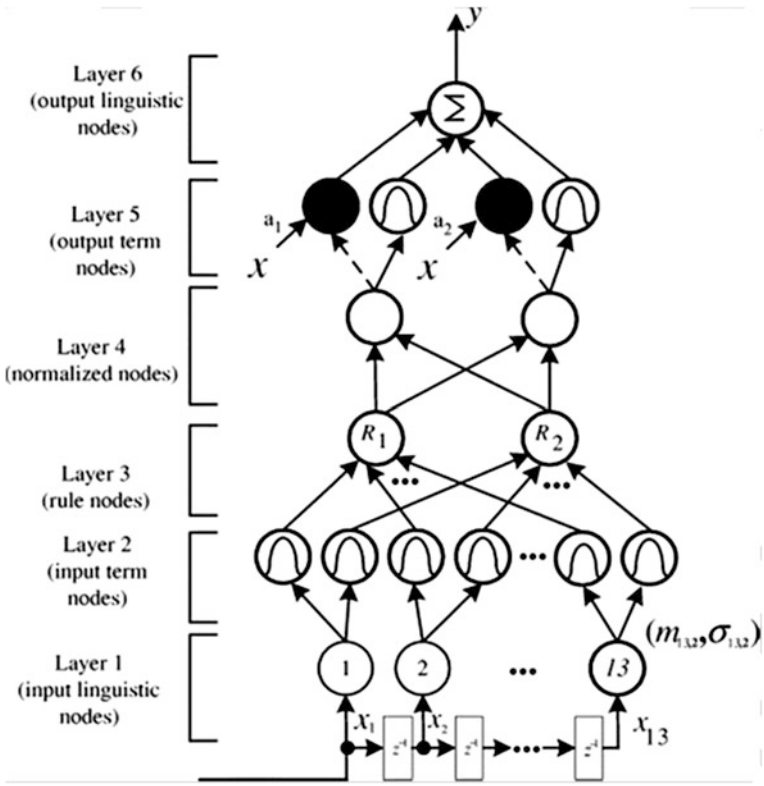


Fig. 5 Structure of the SONFIN for pulse compression by three-element Barker code [12]

parameter. The SONFIN consists of nodes, each of which has some finite fan-in of connections represented by weight values from other nodes and fan-out of connections to other nodes. Associated with the fan-in of a node is an integration function f that serves to combine information, activation, or evidence from other nodes. This function providing the net input for the node is, $a^{(k)} = f\left(u_1^{(k)}, u_2^{(k)} \ldots\ldots u_p^{(k)}; w_1^{(k)}, w_2^{(k)} \ldots\ldots w_p^{(k)}\right)$, where $u_1^{(k)}, u_2^{(k)} \ldots\ldots u_p^{(k)}$ are inputs to this node and $w_1^{(k)}, w_2^{(k)} \ldots\ldots w_p^{(k)}$ are the associated link weights, and $a^{(k)}$denotes the activation function. The superscript (k) in the above equation indicates the layer number.

We shall describe the functions of the nodes in each of the six layers of the SONFIN as follows. Each node in *Layer-1* corresponds to one input variable and only transmits input values to the next layer directly, i.e. $a^{(1)} = u_i^{(1)}$. In *Layer-2*, each node corresponds to a linguistic label (small, large, etc.) of one of the input variables in *Layer-1*. We choose Gaussian membership function to specify the degree to which an input value belongs to a fuzzy set. The operation performed in this layer is, $a^{(2)} = \exp\left(-\left(u_i^{(2)} - m_{ij}\right)^2 / \sigma_{ij}^2\right)$ where m_{ij} and σ_{ij} are, respectively, the center (or mean) and the width (or variance) of the Gaussian membership function of the j^{th} partition for the i^{th} input variable u_i. Hence, the link weight in this layer can be interpreted as m_{ij}. To represent the firing strength of the corresponding fuzzy rule, we use the nodes of *Layer-3* to represent fuzzy logic rules and perform precondition matching of rules. These nodes are combined by AND operation and expressed as $a^{(3)} = \prod_{i=1}^{q} u_i^{(2)}$, where q is the number of *Layer-2* nodes participating in the IF part of the rule.

Layer-4 is used to normalize the firing strength and expressed as $a^{(4)} = u_i^{(4)} / \sum_{i=1}^{r} u_i^{(4)}$, where '*r*' is the number of rule nodes in *Layer-3*. The consequent output is calculated in *Layer-5*. The input variables plus a constant construct the linear combination of the node operation. Thus, the whole function performed by this layer is $a^{(5)} = \left(\sum_j b_j^i x_j + m_0^i\right) u_i^{(5)}$. Finally, the node of *Layer-6* integrates all the actions recommended by *Layer-5* and acts as a defuzzifier with the expression of $a^{(6)} = \sum_{i=1}^{t} u_i^{(6)}$, where t is the number of nodes in *Layer-5*. Two types of learning, structure and parameter learning, are used concurrently for constructing the SONFIN. A detailed description of the overall learning algorithms can be found in [33].

The SONFIN is repeatedly trained offline with the training set being composed of the 26 time-shifted sequences of the B13 code [12]. The training data are generated by simulating the received sequence of a true B13 code as well as a {0} sequence that represents radar has not received any information yet. In these training sequences, the desired output of the SONFIN, y_d is 1 when the proper Barker code just presenting in the input and is 0 otherwise.

5 Comparison of Golay Code and Neuro-Fuzzy Based Pulse Compression Techniques

In the following section, we have compared the GCPC and the NFPC techniques. The basis of this comparison is the computational complexities/requirements of these algorithms and the characteristics of the output of these algorithms i.e. nature of the results of the pulse compression. The main criteria for this comparison is the range side lobe suppression, computational complexity, noise rejection capability, range resolution capability, Doppler tolerance capability, detection performance, maximum ranging capability and the training and convergence requirements of these pulse compression algorithms. Here, we discuss these comparison criteria in detail.

5.1 Range Side Lobe levels

Both of these types of pulse compression techniques are developed for target detection along with the suppression of range side lobes that appear during the pulse compression process. GCPC gives zero side lobes at the output as discussed in Sect. 3. Complementary Golay codes based phase coded signals are transmitted in a specific direction expecting the presence of target(s). After reflection from the target(s) these signals arrive with some delay at the receiver [1–11].

At the receiver correlation of each Golay code is carried out with itself at baseband. On the addition of these two auto-correlation results, the side lobes in the two correlation results being out of phase to each others are canceled out. The final correlation result has only the main lobe present, and is free from side lobes [34–39]. We demonstrate this side lobe suppression in the pulse compression of a Golay code pair of length 64. Auto correlation of one member (code-C) of the Golay code pair of length 64 is given in Fig. 6. This auto-correlation has range side lobes present. Auto correlation of second member (code-K) of the Golay code pair is carried out separately. When the two correlation results are added, the side lobes are completely suppressed as shown in Fig. 7 and have zero side lobe levels.

On the other hand, the Neuro-Fuzzy based techniques do not achieve the zero side lobe levels [12, 31–33]. In neural network based pulse compression techniques, the range side lobes are reduced but not completely suppressed [12, 31–33]. For error free target detection, the range side lobes must have zero levels [1–11].

Due to the presence of range side lobes, false alarms may occur or weak targets may be masked in the side lobes [1–11]. So, ideally zero level range side lobes are required in ranging applications especially in the detection and ranging of multiple and weak targets [1–11]. Due to the presence of range side lobes, the range resolution capability of the ranging system is also impeded i.e. it cannot detect and resolve the very closely spaced targets due to the presence of range side lobes [1–11].

Golay code based pulse compression is seen to be superior to the Neuro-Fuzzy based pulse compression techniques for the range side lobe suppression comparison

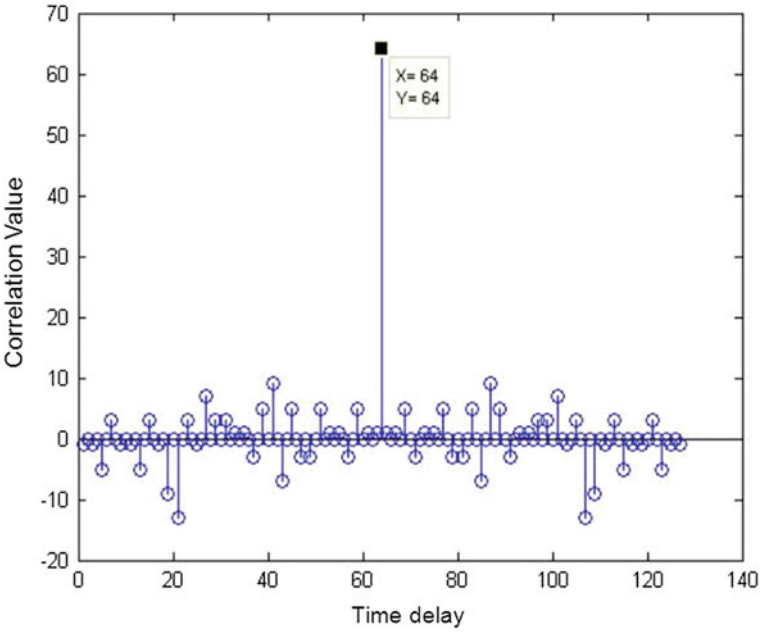


Fig. 6 Auto-correlation of code-C of Golay code pair

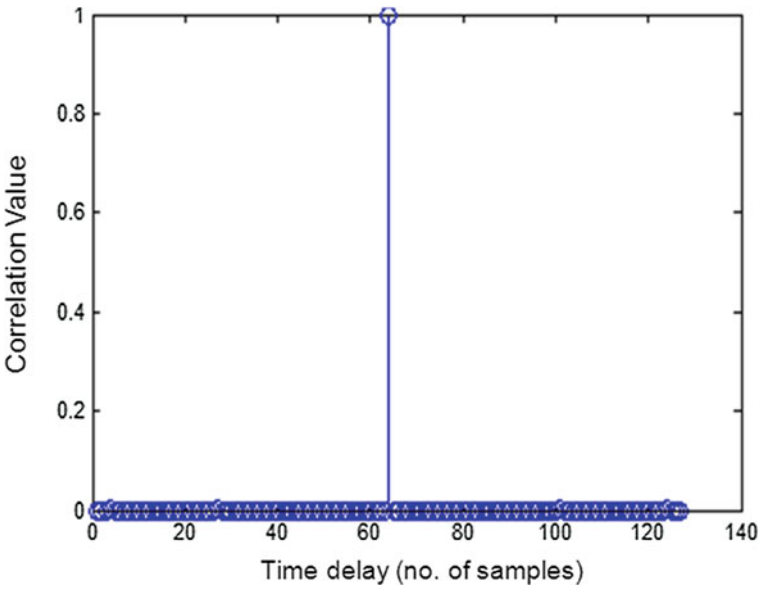


Fig. 7 Suppressed side lobes in the final correlation result of Golay code

criterion. Zero side lobe level in the correlation of Golay codes avoids the disadvantages of minor side lobe levels present in the compressed pulse result of the neural fuzzy pulse compression techniques. These side lobes are a major concern in the detection and ranging of multiple closely spaced and weak targets.

5.2 Computational Complexity

Our second comparison criterion of the two pulse compression techniques is the computational complexity of these techniques. Computational complexity of an algorithm or technique defines how much resources and time is required for the implementation of that technique. Golay code based and the Neuro-Fuzzy based pulse compression techniques both have high computational complexity.

In GCPC, the correlation process is carried out twice and then the correlation results are added to get range side lobe suppression [34–39]. On the other hand, the neural network based pulse compression techniques are computationally very intense [12, 31–33]. They involve online learning and training which requires many resources [12, 31–33]. The details of computational complexity of neural network based pulse compression are already discussed in Sect. 4. Also NFPC techniques are multi layered algorithms which are computationally much intense [12, 31–33].

5.3 Noise Rejection Capability

In the presence of random noise, side lobes appear in the correlation/compressed pulse result [1–11]. Peak of the compressed pulse result can be distinguished by putting a threshold on the noise level. For our analysis of the noise rejection capability of these pulse compression techniques under consideration, additive white Gaussian noise of zero mean and having certain variance is used. Different values of noise variance are used to get different Signal to noise ratio (SNR) values [42–45]. For example, $X = awgn\ (x;10)$ adds an additive white Gaussian noise into a signal "x" to make the signal to noise ratio (SNR) of the resultant signal (X) equal to 10. Then correlation of the white noise corrupted signal is carried out with the clean sequence or pulse compression processing is carried out in the NFPC algorithms.

In case of Golay code, the side lobes are suppressed in the addition of the autocorrelation results of the two complementary codes as shown in Fig. 7. For Golay code, the correlation in case of noise corrupted signal is shown in Fig. 8. The presence of noise is seen along with the main lobe in the final pulse compression result.

Simulations were carried out to investigate the noise rejection capability of both of these types of pulse compression techniques. For the neural network based pulse compression techniques, the input signals used to evaluate the noise robustness are generated by a B13 code, and a 20-element combined Barker codes (CBC), and

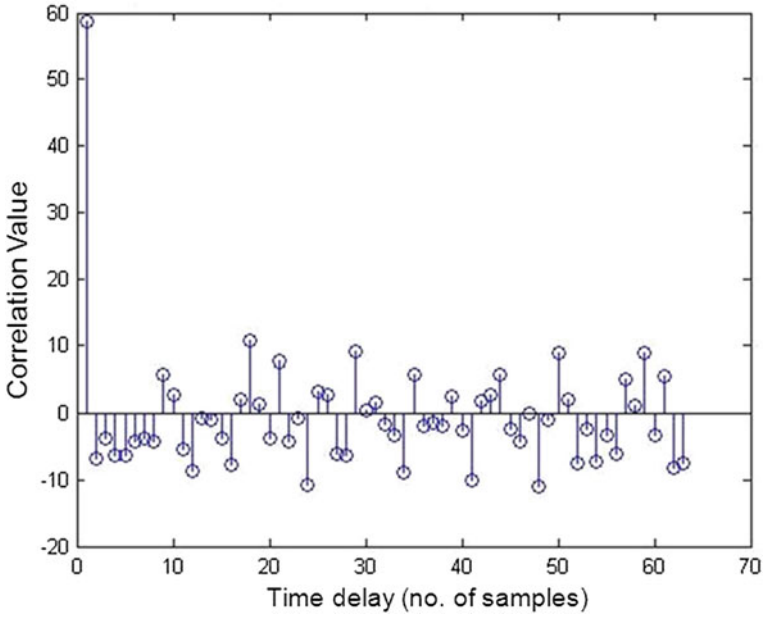


Fig. 8 Correlation of Golay codes in the presence of additive white Gaussian noise

both of them are perturbed by additive white Gaussian noise with five different noise strengths, respectively [13, 31–33].

Effect of the noise variance on the pulse compression output results (main lobe and the side lobes) of the noise corrupted versions of these coding signals were investigated and results are presented in Table 1. The Signal to Side lobe Ratio (SSR in dB) of the compressed pulse signal and its Integrated Side Lobe (ISL in dB) results (Table 1) show that Golay code based pulse compression is robust to noise than the NFPC. Golay code based pulse compression has rejected noise more than the neural fuzzy pulse compression techniques. This noise rejection is highly desired in ranging applications. False alarms may occur and the weak targets may be masked in the noise.

5.4 Range Resolution Capability

The Ability to resolve two closely separated targets in range is called range resolution [1, 2, 5–9]. Range resolution ability is the examination of the ability to distinguish between two closely spaced targets solely by measurement of their ranges in a radar system. Range resolution is directly proportional to pulse width.

$$\text{Range Resolution} = (\text{C x PW})/2 \text{ (meters)} \tag{5}$$

where 'PW' is pulse width and C is the speed of light.

Table 1 Comparison of the noise rejection capability

Algorithm	Signal main lobe to side lobe ratio, [ISL] in dB				
	$\sigma_n = 0.0$	$\sigma_n = 0.1$	$\sigma_n = 0.3$	$\sigma_n = 0.5$	$\sigma_n = 0.7$
GCPC	60.75, −59.80	60.45, −59.10	58.55, −55.25	54.75, −51.10	52.85, −47.43
NFPC	59.45, −56.95	59.19, −56.77	56.62, −52.26	52.41, −48.15	50.58, −44.88

To resolve two targets in range, the basic criterion is that the targets must be separated by at least the range equivalent of the width of the processed echo pulse [1–11]. Coding of a pulse allows a radar to utilize a long pulse to achieve large radiated energy, but simultaneously to obtain the range resolution of a short pulse [1–11]. Bi-phase coding produces the greatest band width desirable for range resolution [1, 2, 5–7]. In other words it provides the minimum possible range apart at which two targets can be detected and resolved. In case of phase coding [1–11]

$$\text{Range Resolution} = (\text{C x Chip width})/2. \tag{6}$$

In Fig. 9, an un-coded pulse is shown with its pulse compression/correlation result. Also its range resolution capability is demonstrated by showing the detection of the nearest possible detectable second target.

In Fig. 10, a bi-phase coded pulse is shown along with its pulse compression/correlation result. Also, its range resolution capability is demonstrated by showing the detection of the nearest possible detectable second target. It can be seen that by phase coding the pulse, its range resolution capability is improved much. Bandwidth of the Phase coded pulse is given by (relation B = ...) [1–11]. Its pulse compression ratio is given by (relation m = ...) [1–11].

$$B = 1/\tau_n^{\cdot} \quad m = \tau/\tau_n = B\tau$$

With the fixed pulse amplitude, the pulse length defines the max power to be transmitted and thus defines the maximum detectable range [1–11]. The developed neural networked pulse compression techniques use Barker codes as the phase coding signals. Barker codes are short in length. Maximum length of the known Barker code is 24. Hence, to have a longer detection range the pulse will be of longer length which means the sub-pulse will also be very wide [1–11].

From the RADAR literature, we know that the range resolution of a phase coded pulsed signal is directly proportional to the sub-pulse width or the chip width of the coding signal [1–11]. Golay codes exist in all lengths [35–39]. Hence, it results that the Golay code based pulse compression can be used for ranging of the targets with very long ranges; and simultaneously it can provide very good range resolution. In other words, it can still detect and resolve very closely spaced targets at very longer ranges.

On the other hand, long Barker codes do not exist [1–11]. Its chip width in a longer pulse will be very wide which reduces the range resolution capability of the NFPC based ranging system. Hence, to attain a very good range resolution is not possible with NFPC while detecting the targets at very longer ranges.

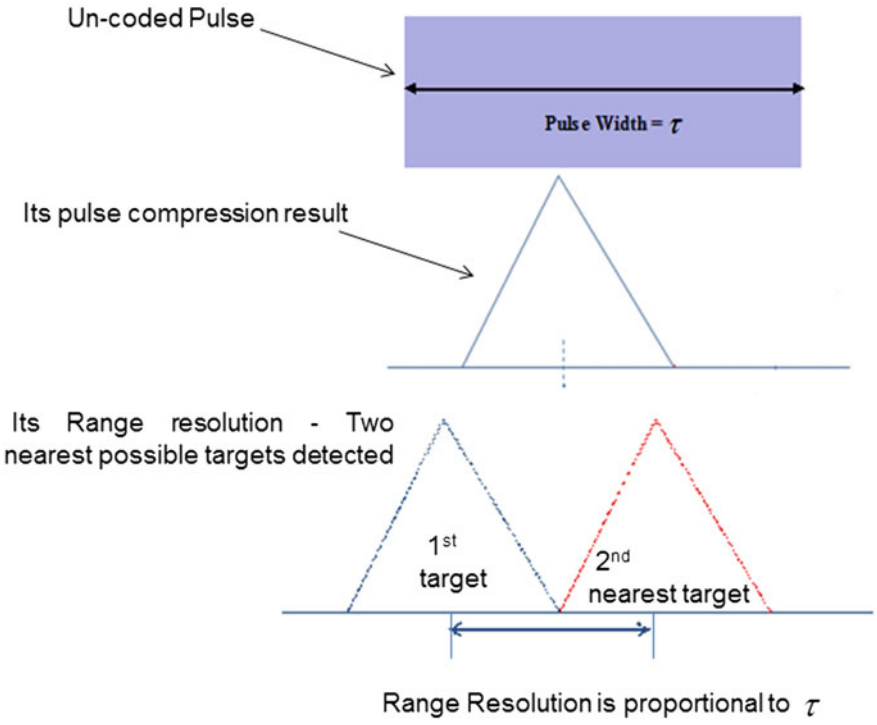


Fig. 9 An un-coded pulse, its pulse compression/correlation result and range resolution capability demonstration

5.5 Doppler Tolerance Capability

When a transmitted pulse is reflected from a moving object, a Doppler shift is introduced in the frequency of the pulse [1–11]. If the Doppler shift is not correctly estimated at the receiver, correlation result may deteriorate and this could cause errors in detection [1–11]. Let the Golay code c-coded pulse c[n] be transmitted. When this pulse is reflected from a moving target, a Doppler shift and some phase is introduced in the received signal. The resultant signal, after preprocessing, is given as [42,44,45]

$$s[n] = c[n]e^{(-jw_D n+\theta)}. \tag{7}$$

The Fourier transform pair of the resultant signal is [42, 44, 45]

$$c[n]e^{(-jw_D n)} \Leftrightarrow C(j(w + w_D)). \tag{8}$$

The spectrum of $s[n]$ is shifted by a frequency of ω_D [42]. The correlation of the transmitted coded pulse with the received signal is given by [1–11, 42, 44, 45]

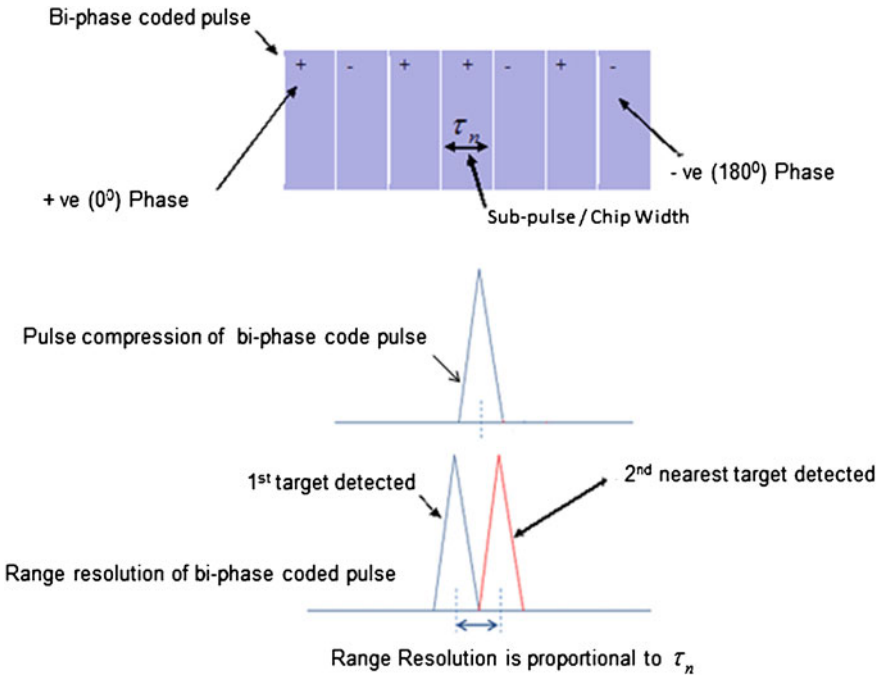


Fig. 10 A bi-phase coded pulse, its compressed pulse result and its range resolution capability demonstration

$$r[n] = \sum_{n}^{N-1} c[k - n_0] e^{(-j(w_D + k))} c[n - k]. \tag{9}$$

When k = n,

$$r[n] = e^{-j\theta} \sum_{n}^{N-1} e^{-jw_D}. \tag{10}$$

The mathematics shows that as Doppler frequency increases, correlation will decrease. Simulations show this effect very clearly and also show the correlator to act as a low pass filter for Doppler shift.

Detection of a single stationary target at a certain range when Golay code is the coding signal is shown in Fig. 11. In the presence of a Doppler frequency the peak of the compressed pulse result reduces as shown in Fig. 12. The side lobes along the frequency axis are due to the non-orthogonal Doppler frequency [42–45]. The difference in the correlation level in Figs. 11, 12 is evident.

We have made a comparative study of the effect of increase in Doppler frequency on the peak of the compressed pulse result obtained from both of the techniques. Equal Doppler shifts were introduced into the input signals of both

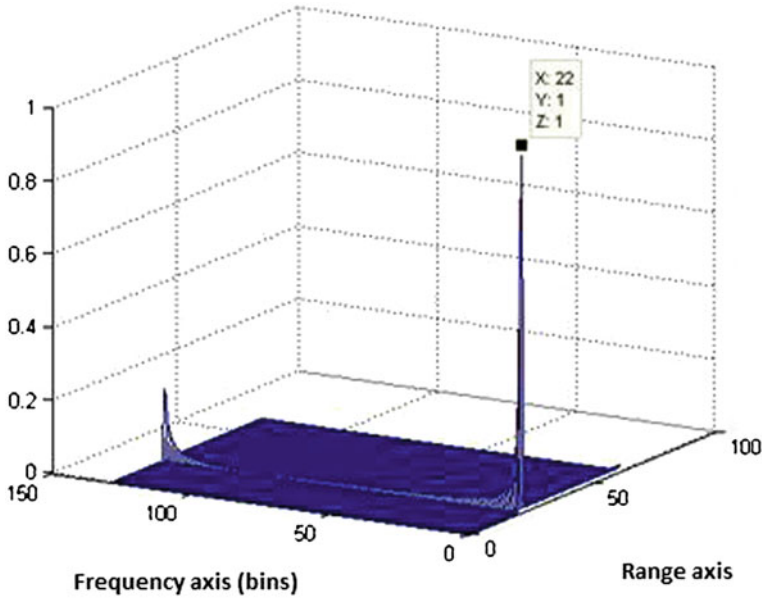


Fig. 11 Detection of a single stationary target using Golay code

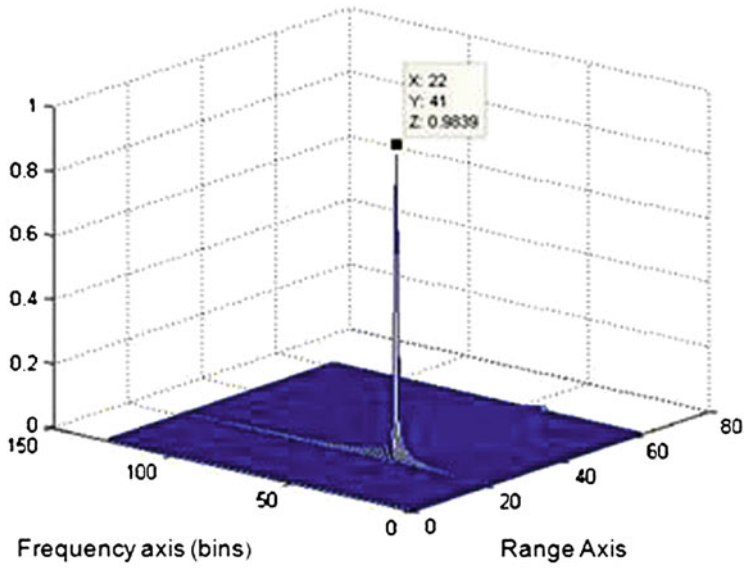


Fig. 12 Reduced correlation peak in the detection of a moving target at a certain range

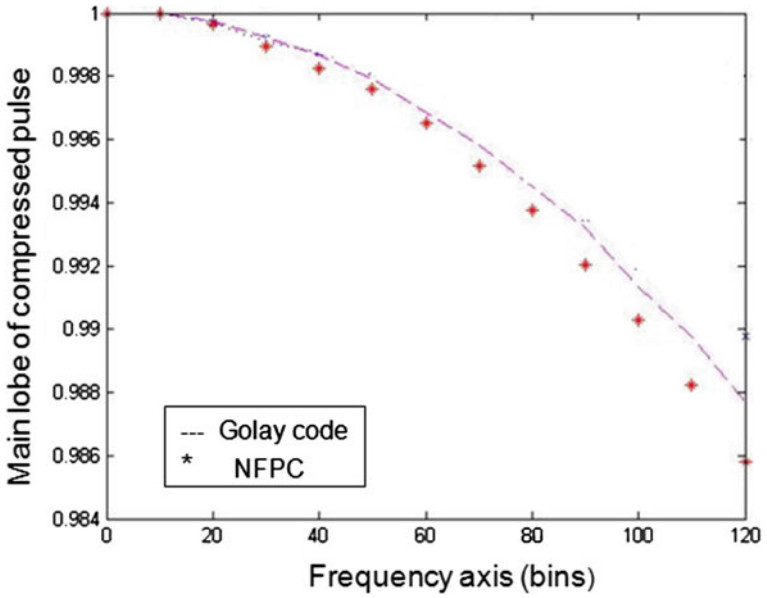


Fig. 13 Effect of increase in Doppler shift on main lobe of compressed pulse

the pulse compression techniques and the reduction in the peak of the main lobe of the compressed pulse was recorded and plotted in Fig. 13. Simulation results show that peak of the main lobe of the compressed pulse result reduces in the presence of the Doppler shift. When the Doppler shift increases, the peak of the compressed pulse reduces.

This comparison shows that the Golay code based pulse compression is less affected by the Doppler shift. It shows the Doppler tolerance superiority of Golay code based pulse compression technique over the NFPC.

5.6 Training and Convergence Requirements of the Algorithm

Golay code based pulse compression does not require any training and convergence requirements [34–39]. However, the neural network and fuzzy based pulse compression algorithms have the training and convergence requirements [12, 31–33]. The details of the training and convergence requirements of NFPC techniques are given in [12, 31–33]. Extra time and more computational efforts are required for this [12, 31–33]. Also, the convergence speed of most of the neural networked and Fuzzy algorithms is inherently low [12–31].

6 Conclusion

We have made a comparison of the Golay code based pulse compression (GCPC) and the neural network based pulse compression techniques for ranging applications. Both of these pulse compression techniques are developed for target detection in pulsed ranging systems and simultaneously providing range side-lobe suppression. Among the neural network based pulse compression techniques, NFPC technique is selected for this comparison due to its fast training and convergence speed, high signal-to-side lobe ratio than the traditional neural network based pulse compression algorithms and low integrated side lobe levels. The criteria for the comparison of these pulse compression techniques are based on the analysis of the output of these pulse compression techniques and their computational requirements. These comparison criteria include the range side-lobe suppression, computational complexity, noise rejection, Doppler tolerance, range resolution capability and the training and convergence requirements of these pulse compression techniques. GCPC provides zero level range side lobes while range side lobes are reduced considerably but not completely suppressed in the NFPC. Range side lobe suppression is a very critical performance measure as the presence of range side lobes may lead to false alarms and weak targets may be masked in them. This phenomenon is very important in the detection and ranging of multiple closely spaced and weak targets. GCPC based pulse compression is robust to noise than the NFPC. Noise rejection is also a very critical performance measure of a ranging algorithm. The presence of noise in the compressed pulse result has the same implications as of the range side lobes. Both of these pulse compression techniques are computationally intensive. NFPC is computationally more intense as it has multiple layers of computations. More time and computational resources are required for NFPC. Also, it has training and convergence requirements which require more time and resources. GCPC does not require training and convergence and it can provide better range resolution while detecting the targets at very longer ranges. NFPC is based on short length Barker codes and cannot detect and resolve very closely spaced targets located at very longer ranges. GCPC is less affected by the Doppler shift corresponding to the target speed than NFPC. The presence of the large Doppler shift reduces the level of the main lobe of the compressed pulse result which may laid to a wrong detection decision. On average, GCPC has shown better performance in our comparison. However, to use complementary code Golay pair as the phase coding signal requires some special considerations. Also, it may lead to range ambiguity problem if transmitted on separate pulses. This problem can be solved if Golay code pair is transmitted on a bi-carrier pulse. NFPC has no range ambiguity problems, which is a clear advantage over GCPC.

References

1. D.K. Barton, *Modern Radar System Analysis* (Artech House Inc., Norwood, 2008)
2. M. Bassem, *Radar System Analysis and Design using Matlab* (Chapman and Hall/CRC, New York, 2000)
3. S.M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, vol 1 (Prentice-Hall, Upper Saddle River, 1993)
4. S.M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*, vol 2 (Prentice-Hall, Upper Saddle River, 1998)
5. M.A. Richards, *Fundamentals of Radar Signal Processing* (MC Graw-Hill, New York, 2005)
6. F.E. Nathanson, *Radar System Fundamentals: Signal Processing and the Environment* (SciTech Publishing Inc., NJ, 1999)
7. N. Levanon, E. Mozeson, *Radar Signals* (John Wiley and Sons, NJ, 2004)
8. M.H. Hayes, *Statistical Digital Signal Processing and Modeling* (John Wiley and Sons (Asia) Pvt Ltd, Singapore, 2003)
9. M. Bassem, *Radar System Analysis and Design using Matlab* (Chapman and Hall/CRC, New York, 2000)
10. M.I. Skolnik, *Introduction to Radar Systems*, 2nd edn. (McGraw Hill, New York, 1980)
11. M. Barkat, *Signal Detection and Estimation*, 2nd edn. (Artech House Inc, NJ, 2005)
12. F.B. Duh, C.F. Juang, A neural fuzzy network approach to radar pulse compression. IEEE Geosci. Remote Sens. Lett. **1**(1),15–20 (2004)
13. M.H. Ackroyd, Optimum mismatched filters for side lobe suppression, IEEE Trans. Aerosp. Electron. Syst. AES-**9**, 214–218 (1993)
14. R.C. Daniels, V. Gregers, Code inverse filtering for complete side-lobe removal in binary pulse compression systems, in *Proceedings of the IEEE International Radar Conference 2005*, pp. 256–261
15. R.J. Keller, S.A. Mudukutore, V. Chandrasekhar, Pulse compression for weather radars: simulation and evaluation. IEEE Trans. Geosci. Remote Sens. **36**(1), 125–142 (1998)
16. N.J. Bucci, *Doppler Tolerant Range Side-Lobe Suppression for Pulse Compression Radars, Tech Reference MTMR-TR-001* (GE Aerospace Moorestown, NJ, 1991)
17. K.J. Gartz, Generation of uniform amplitude complex code sets with low correlation side lobes, IEEE Trans. Signal Process. **40**(2), 343–351(1992)
18. R.S. Masanori, Time side lobe reduction technique for binary phase coded pulse compression, in *Proceedings of the IEEE International Radar Conference 2000,* Alexendaria, VA, USA (2000), pp. 809–814
19. X. Wu, S.J.F. Liu, W. Zhao, Chaotic phase code for radar pulse compression, in *Proceedings of the IEEE Radar Conference 2001* (2001), pp. 279–283
20. D.W. Jin, J. Kai, W. Wei-dong, The study of new radar wave form with high range resolution, in *Proceedings of the IEEE International Conference on Signal Processing, (ICSP 2004)* (2004), pp. 1957–1960
21. J. George, N. Bharadivy, V. Chandrasekhar, Waveform coding for dual polarization weather radars, in *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, (IGARSS 2007)* (2007), pp. 3571–3574
22. A. Kusk, J. Dall, Azimuth phase coding for range ambiguity suppression in SAR, in *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2004)*, pp. 1734–1737
23. H. Shen, W. Zhang, S.K. Kwak, The modified chirp UWB ranging system for vehicular applications, in *Proceedings of the IEEE International Symposium on Communications and Information Technologies, ISCIT* (2007), pp. 1038–1042
24. J.D. Jenshak, J.M. Stils, A fast method for designing optimal transmit codes for radar, in *Proceedings of the IEEE Radar Conference (RADAR 2008)* (2008), pp. 1–6
25. S.R.J. Anelson, Random noise radar/sonar with ultra wideband waveforms, IEEE Trans. Geosci. Remote Sens. **45**(5), 1099-1114 (2007)

26. S.R.J. Anelson, Suppression of noise floor and dominant reflectors in random noise radar, in *Proceedings of the International Radar Symposium 2006* (*IRS 2006)* pp. 1–6
27. R.J. Keeler, C.A. Hwang, Pulse compression for weather radar, in *Proceedings of the IEEE International Radar Conference* (May 1995) pp. 529–535
28. J.M. Ashe, R.L. Nevin, D.J. Murrow, H. Urkowitz, N.J. Bucci, J.D. Nespor, Range sidelobe suppression of expanded/compressed pulses with droop, in *Proceedings of the 1994 IEEE International Radar Conference,* Atlanta, GA, 29–31 March 1994, pp. 116–122
29. C.A. Hwang R.J. Keeler, Sample phase aspects of FM pulse compression waveforms, in *Proceedings of the IGARSS* (1995), pp. 2126–2128
30. R.J. Keeler, C.A. Hwang, Pulse compression for weather radar, in *Proceedings of the IEEE International Radar Conference* (May 1995), pp. 529–535
31. N.J. Bucci, H.S. Owen, K.A. Woodward, C.M. Hawes, Validation of pulse compression techniques for meteorological functions, IEEE Trans. Geosci. Remote Sens. **35**, 507–523 (1997)
32. H. K. Kwan and C. K. Lee, A neural network approach to pulse radar detection, IEEE Trans. Aerosp. Electron. Syst. **29**, 9–21 (1993)
33. K.D. Rao G. Sridhar, Improving performance in pulse radar detection using neural networks, IEEE Trans. Aerosp. Electron. Syst. **31**, 1194–1198 (1995)
34. C.F. Juang C.T. Lin, An on-line self-constructing neural fuzzy inference network and its applications, IEEE Trans. Fuzzy Syst. **6**, 12–32 (1998)
35. M.J. Golay, Complementary series, IRE Trans. Inf. Theory 7, 82–87 (1961)
36. C. Hackett, Correction to An efficient algorithm for soft-decision decoding of the (24, 12) extended Golay code. IEEE Trans. Commun. **234**(43), 812–816 (1995)
37. J. Snyders, M. Ran, Constraint designs for maximum likelihood soft decoding of RM (2, M) and the extended Golay codes. IEEE Trans. Commun. **43**(234), 169–176 (1995)
38. B. Hodgins, K. Harman, The next generation of Guidar Technology, in *Proceedings of the IEEE Security Technology 2004, 38th Annual 2004 International Carnhan Conference* (2004), pp. 169–176
39. S. Searle, S. Howard, A novel poly-phase code for side-lobe suppression, in *Proceedings of the IEEE Waveform Diversity and Design 2007* (2007), pp. 377–381
40. R.Y. Chio, X. Hao, Coded excitation for diagnostics ultrasound: a system developer's perspective, in *Proceedings of 2003 IEEE Ultrasonics Symposium* (2003), pp. 437–448
41. A. Hussain, M. Rais, M.B. Malik, Golay Codes in ranging applications, in *Proceedings of the 8th IASTED Conference on Wireless and Optical Communications* (*WOC-2008),* Quebec city, Canada (2008), pp. 184–188
42. S. Haykin, *Communication Systems*, 3rd edn. (John Wiley and sons, New York, 1999)
43. A. Hussain, M.B. Malik, Golay Code based carrier phase recovery, in *Proceedings of the 4th IEEE International Conference on Computer and Emerging Technologies (ICET 2008)*, Islamabad, Pakistan (2008) pp. 92–97
44. A.V. Oppenheim, A.S. Wilsky, S.H. Nawab, *Signals and Systems*, (Prentice-Hall, Upper Saddle River, 1993)
45. A.V. Oppenheim, R.W. Shaffer, S.H. Nawab, *Discrete Time Signal Processing*, 2nd edn. (Prentice-Hall, Upper Saddle River, 1999)

Part III
Robotics

Trajectory Generation for Autonomous Mobile Robots

Vu Trieu Minh

Abstract This chapter presents the generation of car-like autonomous mobile robots/vehicles tracking trajectory with three different methods comprising of flatness, polynomial and symmetric polynomial equations subject to constraints. Kinematic models for each method are presented with all necessary controlled variables including position, body angle, steer angle and their velocities. The control systems for this model are designed based on fuzzy/neural networks. Simulations are analyzed and compared for each method. Studies of this chapter can be used to develop a real-time control system for auto-driving and/or auto-parking vehicles.

Keywords Trajectory generation · Autonomous mobile robot · Nonholonomic · Flatness · Polynomial

1 Introduction

This chapter studies the problem associated with trajectory generation for car-like autonomous mobile robots moving from an initial point to any final point subject to constraints. This study can be used to develop a real-time control system for autonomous ground vehicles which can track on any feasible paths from the global positioning system (GPS) maps or/and from unmanned aerial vehicle (UAV) images. This system can be applied to autonomous unmanned ground vehicles (on road or off road), and auto-parking, auto-driving systems. The vehicles can perform intelligent motion without requiring a guidance command or remote teleoperation.The main idea of this chapter is to propose a system which can

V. T. Minh (✉)
Mechanosystem—Department of Mechatronics, Tallinn University of Technology, Tallinn, Estonia
e-mail: vutrieuminh@yahoo.com

Z. H. Khan et al. (eds.), *Computational Intelligence for Decision Support in Cyber-Physical Systems*, Studies in Computational Intelligence 540,
DOI: 10.1007/978-981-4585-36-1_6,

automatically generate an optimally feasible trajectory, and then, control the autonomous vehicles/robots to track exactly on this path from any given starting point to any desired destination point from a map subject to the vehicle's physical constraints due to obstacles, speed, steer angle, etc.

In order to deal with the constrained trajectory generation problem, all constraints from a real vehicle must be included into the equations such as the size, position, body angle, steer angle and their velocities. Looking at the current research articles on this topic, there are very few of them dealing with real automotive engineering constraints. Basic introduction on flatness, nonholonomic, and nonlinear systems can be read from Levine [7], where the fundamental motion planning of a vehicle is presented. Parking simulation of 2-trailer vehicle is also demonstrated but without constraints of steer angle and steer angular velocity. The problem of trajectory generation for nonholonomic system is also well considered by Dong and Guo [3], where two trajectory generation methods are proposed. The control inputs are the second order polynomial equations. By integrating those control inputs, coefficients for the second order polynomial equations are found. However, this article is lacking constraints analysis on the vehicle velocity as well as the steer angle.

Optimal control based on cell mapping techniques for a car-like robot is studied by Gomez [4] subject to the energy-optimal constraint and based on bang-bang control theory. This article shows a simulation of a wheeled mobile robot moving on a path with the steering angle velocity control. However, the article does not mention the algorithms for generating the vehicle trajectory. Several other research articles on optimal trajectories and control of autonomous mobile robots can be seen e.g. by Wang et al. [13], Werling et al. [14], Kanjanawanishkul et al. [5], and Klancar [6]. However, most of those studies are based on the real traffic flow and the control algorithms are to perform the maneuver tasks such as lane-changing, merging, distance-keeping, velocity-keeping, stopping and collision avoidance etc.

This book chapter, therefore, concentrates on the applicable mathematic algorithms to generate optimal trajectories from any start point to any destination point subject to feasible vehicle constraints. This book chapter is the continuation of the previous research article on vehicle sideslip model and estimation by Minh [8]. Nonlinear computational schemes for the nonlinear systems are discussed from Minh and Nitin [11] and Minh and Fakhruldin [10]. Although, the holonomous-dynamical trajectories generation is not new, recent research articles have tried to develop those trajectories in different applications: Trajectory for autonomous ground vehicles based on Bezier curves operating under waypoints and corridor constraints can be referred to Minh [9] and Choi et al. [1]. A real-time trajectory generation for car-like vehicles navigating dynamic environments is discussed in Delsart et al. [2]. A recent paper on trajectory generation of model-based IMM tracking for safe driving in intersection scenario is described in Zhou et al. [15] to reduce traffic accidents at intersections. Another related work on dynamic trajectory generation for wheeled robots is discussed in Missura and Behnke [12] where a dynamic motion of wheeled robots can be determined in real-time onboard.

This chapter introduces three conventional holonomous trajectories generation algorithms (flatness, polynomial and symmetric) subject to steering angle constraint for autonomous mobile robots. Performances of these three algorithms are compared. Data taken from the simulation of the vehicle's longitudinal velocity and its steering angular velocity will be used to develop the control algorithms for the mobile robot tracking on those trajectories in the next part of this chapter. The outline of this chapter is as follows: Sect. 2 describes the kinematic model for a mobile robot; Sect. 3 presents the flatness method for the mobile robot trajectory generation; Sect. 4 discusses the trajectory generation subject to steer angle constraint; Sect. 5 introduces the polynomial method for trajectory generation; Sect. 6 develops the symmetric polynomial method; Sect. 7 analyses the performances of the three methods and finally the conclusion is drawn in Sect. 8.

2 Kinematic Model of a Mobile Robot

The constraints for the mobile robot model are based on the assumption that the wheels are rolling without slipping and the steering angle is simplified as a single wheel at the midpoint of the two front wheels. Then, a kinematic model of the mobile robot shown in Fig. 1 can be drawn as under.

The kinematic model of a forward rear-wheel driving mobile robotis written as:

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{\theta} \\ \dot{\varphi} \end{bmatrix} = \begin{bmatrix} \cos\theta \\ \sin\theta \\ \frac{\tan\varphi}{l} \\ 0 \end{bmatrix} r v_1 + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} v_2 \tag{1}$$

where $X = [x, y, \theta, \varphi]'$ is the system state variables, (x, y) are the Cartesian coordinates of the middle point of the rear wheel axis, θ is the angle of the vehicle body to the x-axis, φ is the steering angle, l is the vehicle wheel base, r is the wheel radius, v_1 is the angular velocity of the rear wheel, and v_2 is the angular steering velocity. Given the initial state $X(0) = [x_0, y_0, \theta_0, \varphi_0]'$ at time $t = 0$ and the final state $X(T) = [x_T, y_T, \theta_T, \varphi_T]'$ at time $t = T$, the paper generates a feasible trajectory for this vehicle. Similarly, the model for a forward front-wheel driving vehicle is presented as:

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{\theta} \\ \dot{\varphi} \end{bmatrix} = \begin{bmatrix} \cos\theta\cos\varphi \\ \sin\theta\cos\varphi \\ \frac{\tan\varphi}{l} \\ 0 \end{bmatrix} r v_1 + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} v_2 \tag{2}$$

For a vehicle moving in a reverse direction (backing), the velocity, v_1 of this vehicle is assigned a negative value. Then, the general kinematic model in (1) is changed to:

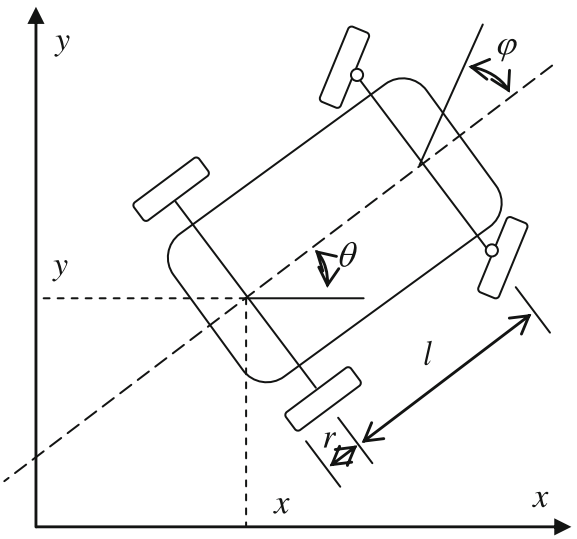


Fig. 1 A simplified mobile robot model

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{\theta} \\ \dot{\varphi} \end{bmatrix} = -\begin{bmatrix} \cos\theta \\ \sin\theta \\ \frac{\tan\varphi}{l} \\ 0 \end{bmatrix} rv_1 + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} v_2 \tag{3}$$

From the above three fundamental equations, flatness equations for the moving vehicle are investigated in the next section.

3 Flatness Trajectory Generation

From Fig. 1, the vehicle angular velocity can be calculated as:

$$rv_1 = \sqrt{\dot{x}^2 + \dot{y}^2} \Rightarrow v_1 = \frac{\sqrt{\dot{x}^2 + \dot{y}^2}}{r} \tag{4}$$

Transformation from Eq. (1), the vehicle body angle is:

$$\frac{\dot{y}}{\dot{x}} = \frac{rv_1 \sin\theta}{rv_1 \cos\theta} = \tan\theta \Rightarrow \theta = \arctan\left(\frac{\dot{y}}{\dot{x}}\right) \tag{5}$$

From the derivative of the above trigonometric relationship for body angular velocity (θ), its derivative $\dot{\theta}$, is achieved as follows:

$$\dot{\theta} = \frac{\ddot{y}\dot{x} - \ddot{x}\dot{y}}{\dot{x}^2} \frac{1}{\left(\frac{\dot{y}}{\dot{x}}\right)^2 + 1} = \frac{\ddot{y}\dot{x} - \ddot{x}\dot{y}}{\dot{x}^2 + \dot{y}^2} = \frac{\tan\ \varphi}{l} r v_1 \Rightarrow \varphi = \arctan\left(l \frac{\ddot{y}\dot{x} - \ddot{x}\dot{y}}{r(\dot{x}^2 + \dot{y}^2)^{\frac{3}{2}}}\right) \tag{6}$$

Therefore, θ and φ can be directly calculated from other variables, i.e. $\dot{x}$, $\ddot{x}$, and $\dot{y}$, $\ddot{y}$. And it means that the above system is flat in Levine [7]. Thus, all state and input variables can be presented by the flat outputs x and y. The boundary conditions for the outputs x and y are:

$$\begin{aligned} \frac{\partial y}{\partial x} = \tan\ \theta \Rightarrow \frac{\partial^2 y}{\partial x} = \dot{\theta} \frac{1}{\cos^2\ \theta} = \frac{\tan\ \varphi}{l} r v_1 \frac{1}{\cos^2\ \theta} = \frac{\tan\ \varphi}{l} \frac{\partial x}{\cos\ \theta} \frac{1}{\cos^2\ \theta} \\ \Rightarrow \frac{\partial^2 y}{\partial x^2} = \frac{\tan\ \varphi}{l \cos^3\ \theta} \end{aligned} \tag{7}$$

For $x(t)$, the initial state at time, $t\ = 0$, to the final state at time, $t\ = T$:

$$x(0) = x_0 \Rightarrow x(T) = x_T \tag{8}$$

The initial state for $y(t) = y(0)$ is given as:

$$y(0) = y_0 = \tan\ \theta_0 \Rightarrow \frac{\partial^2 y}{\partial x^2}|_{t=0} = \frac{\tan\ \varphi_0}{l \cos^3\ \theta_0} \tag{9}$$

and the final state for $y(t) = y(T)$:

$$y(T) = y_T = \tan\ \theta_T \Rightarrow \frac{\partial^2 y}{\partial x^2}|_{t=T} = \frac{\tan\ \varphi_T}{l \cos^3\ \theta_T} \tag{10}$$

From the initial state $(x_0, y_0, \theta_0, \varphi_0)$ at the time $t = 0$ to the final state $(x_T, y_T, \theta_T, \varphi_T)$, and under a real condition that $|\dot{x}(t)| \geq \varepsilon > 0$. If it is assumed that, $\varepsilon = \frac{|x_T - x_0|}{2T} > 0$, the trajectory of $x(t)$ can be selected freely as:

$$x(t) = \left(\frac{T - t}{T}\right) x_0 + \frac{t}{T} x_T + |x_T - x_0| \frac{t(t - T)}{2T^2} \tag{11}$$

And the trajectory of $y(t)$ can be selected as:

$$y(t) = y_0 + t\alpha_1 \tan\ \theta_0 + t^2 \frac{\alpha_2 \tan\ \varphi_0}{2l \cos^3\ \theta_0} + t^3 b_1 + t^4 b_2 + t^5 b_3 \tag{12}$$

where, $\alpha_1 = \frac{2(x_T - x_0) - |x_T - x_0|}{2T}$, $\alpha_2 = \frac{|x_T - x_0|}{T^2}$, and $\alpha_3 = \frac{2(x_T - x_0) + |x_T - x_0|}{2T}$;

and $b = [b_1, b_2, b_3]' = A^{-1}c$ with:

$$A = \begin{bmatrix} T^3 & T^4 & T^5 \\ 3T^2 & 4T^3 & 5T^4 \\ 6T & 12T^2 & 20T^3 \end{bmatrix} \text{ and } c = \begin{bmatrix} y_T - y_0 - T\alpha_1 \tan\ \theta_0 - T^2 \frac{\alpha_2 \tan\ \varphi_0}{2l \cos^3\ \theta_0} \\ \alpha_3 \tan \theta_T - \alpha_1 \tan\ \theta_0 - T \frac{\alpha_2 \tan\ \varphi_0}{l \cos^3\ \theta_0} \\ \frac{\alpha_2 \tan\ \varphi_T}{l \cos^3\ \theta_T} - \frac{\alpha_2 \tan\ \varphi_0}{l \cos^3\ \theta_0} \end{bmatrix}.$$

From Eq. (10),

$$\theta = \arctan\left(\frac{2T^2\left(\alpha_1 \tan\,\theta_0 + \frac{\alpha_2 \tan\,\varphi_0}{l\cos^3\,\theta_0}t + 3b_1t^2 + 4b_2t^3 + 5b_3t^4\right)}{(2T(x_T - x_0) - T|x_T - x_0| + 2|x_T - x_0|t)}\right) \tag{13}$$

and from Eq. (5),

$$\begin{aligned}\varphi &= \arctan\left(\frac{\partial^2(y) l\cos^3\,\theta}{\partial x^2}\right) \\ &= \arctan\left(\frac{\left(\frac{\alpha_2 \tan\,\varphi_0}{l\cos^3\,\theta_0} + 6b_1t + 12b_2t^2 + 20b_3t^3\right)(2T^2)^2 l\cos^3\,\theta}{(2T(x_T - x_0) - T|x_T - x_0| + 2|x_T - x_0|t)^2}\right)\end{aligned} \tag{14}$$

The angular velocity of the vehicle in Eq. (1) can be calculated from (11) and (12) for y:

$$\dot{y}(t) = \left(\alpha_1 \tan\,\theta_0 + \frac{\alpha_2 \tan\,\varphi_0}{l\cos^3\,\theta_0}t + 3b_1t^2 + 4b_2t^3 + 5b_3t^4\right) \tag{15}$$

and then,

$$\ddot{y}(t) = \left(\frac{\alpha_2 \tan\,\varphi_0}{l\cos^3\,\theta_0} + 6b_1t + 12b_2t^2 + 20b_3t^3\right) \tag{16}$$

And for x:

$$\dot{x}(t) = \left(\frac{(2T(x_T - x_0) - T|x_T - x_0| + 2|x_T - x_0|t)}{2T^2}\right) \tag{17}$$

And then,

$$\ddot{x}(t) = \left(\frac{|x_T - x_0|}{T^2}\right) \tag{18}$$

The absolute vehicle velocity can be calculated from Eqs. (15) to (18):

$$v_1(t) = \frac{\sqrt{\dot{x}^2 + \dot{y}^2}}{r} \tag{19}$$

or with another method to calculate $v_1(t)$ is $v_1(t) = \frac{\dot{x}(t)\cos\,\theta}{r} + \frac{\dot{y}(t)\sin\,\theta}{r}$, then:

$$\begin{aligned}v_1(t) = &\left(\frac{(2T(x_T - x_0) - T|x_T - x_0| + 2|x_T - x_0|t)}{2rT^2}\right)\cos\,\theta \\ &+ \left(\frac{\left(\alpha_1 \tan\,\theta_0 + \frac{\alpha_2 \tan\,\varphi_0}{l\cos^3\,\theta_0} + 3b_1t^2 + 4b_2t^3 + 5b_3t^4\right)\sin\,\theta}{r}\right)\end{aligned} \tag{20}$$

To calculate $\dot{\theta}$ from Eq. (6):

$$\dot{\theta} = \frac{\ddot{y}\dot{x} - \ddot{x}\dot{y}}{\dot{x}^2 + \dot{y}^2} = \frac{\tan\,\varphi}{l} r v_1 \tag{21}$$

To calculate $v_2(t) = \dot{\varphi}$, from Eq. (14), if $\varphi = \arctan(X)$, then,

$$\begin{aligned} \dot{\varphi} &= \dot{X}\left(\frac{1}{1+X^2}\right) \text{ with} \\ X &= \left(\frac{\left(\frac{\alpha_2 \tan\,\varphi_0}{l\cos^3\,\theta_0} + 6b_1 t + 12b_2 t^2 + 20b_3 t^3\right)(2T^2)^2 l\cos^3\,\theta}{(2T(x_T - x_0) - T|x_T - x_0| + 2|x_T - x_0|t)^2}\right) \end{aligned} \tag{22}$$

Another method to calculate $v_2(t) = \dot{\varphi}$ is from Eq. (6) is:

$$\dot{\varphi} = \partial\left(l\frac{\ddot{y}\dot{x} - \ddot{x}\dot{y}}{r(\dot{x}^2 + \dot{y}^2)^{\frac{3}{2}}}\right)\left(\frac{1}{1 + \left(l\frac{\ddot{y}\dot{x} - \ddot{x}\dot{y}}{r(\dot{x}^2+\dot{y}^2)^{\frac{3}{2}}}\right)^2}\right) \tag{23}$$

Simulation parameters of a real vehicle are as follows: $l = 2\ m$, $r = 0.25\ m$, $x(0) = [0, 0, 0, 0]'$, $x(T) = \left[10, 10, 0, \frac{\pi}{6}\right]'$, and $T = 100$, the trajectory generation of this vehicle is generated from Eqs. (13) and (14). Vehicle velocity $v_1(t)$ is generated from Eqs. (19) or (20).

Body angle, θ, is from Eq. (13), and the angular velocity of body angle, $\dot{\theta}$, is from Eq. (21). The steering angle, φ, is from Eq. (14). The steering angular velocity, $\dot{\varphi}$, can be directly calculated form Eqs. (22) or (23). Figure 2 shows the coordinate trajectory (x, y) for the vehicle from initial position to the final position, and the velocity of the vehicle along the track. Figure 3 shows the vehicle body angle (θ), and the steering angle (φ), corresponding to the angular velocity ($\dot{\theta}$, $\dot{\varphi}$) of the vehicle.

From Fig. 3, the maximum steer angle for this movement, $\varphi_{\max} = 64°$, and exceed the physical constraint of a real vehicle with $-45° \le \varphi_{\max} \le 45°$. Therefore the above trajectory is not feasible. The problem is that the distance from the start point to the destination point is too short subject to the real vehicle physical constraints. A solution for this method is to lengthen the travelling distance until it meets the vehicle constraint on steer angle. This issue is discussed in the next section.

4 Trajectory Subject to Steer Angle Constraint

Due to a real structure constraint for a vehicle with $-\frac{\pi}{4} \le \varphi \le \frac{\pi}{4}$ or as mentioned above, the maximal angle for a real vehicle normally in a range of $-45° \le \varphi \le 45°$in the following constraint:

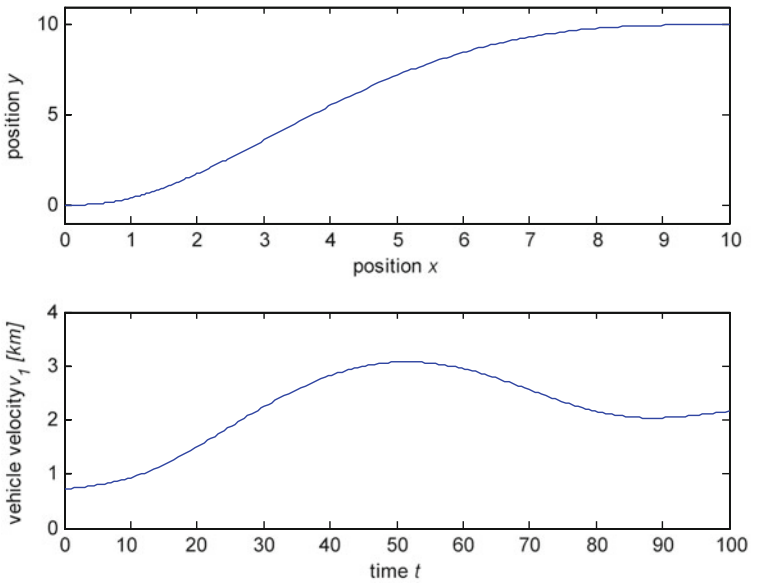


Fig. 2 Trajectory and velocity

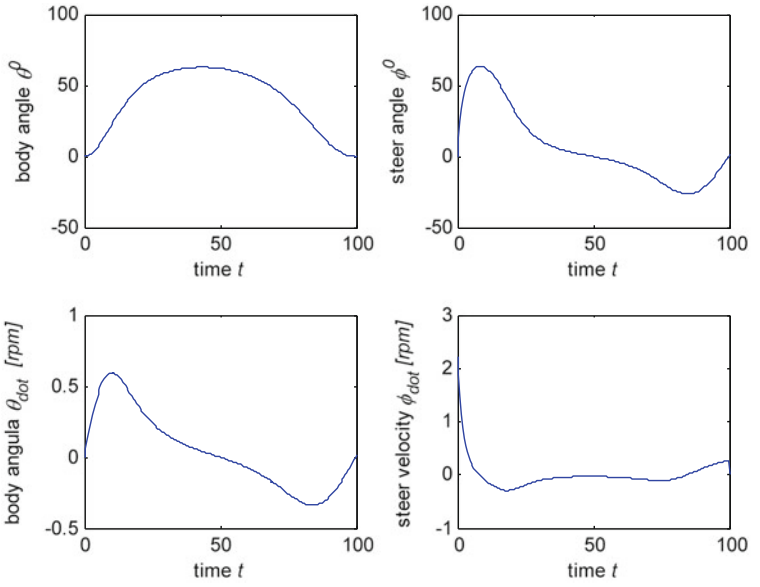


Fig. 3 Body and steer angle

$$-\frac{\pi}{4} \leq \varphi \leq \frac{\pi}{4} \tag{24}$$

Simulation in the previous part shows that the maximal needed steer angle, $\varphi_{\max} = 64^{\circ}$ and exceeds the requirement in (24). Therefore, in this part, a new method for the vehicle trajectory generation subject the constraint in Eq. (24) is proposed.

For time t from initial state, $t = 0$, to the final state, $t = T$, the feasible trajectory $x(t)$ and $y(t)$ subject to conditions in Eqs. (11) and (12) can be generated. For each progress of timing, t, re-calculate the steering angle, φ in Eq. (14), then, check for the steering constraint in Eq. (24). If the constraint in (24) is violated, it means that the distance from the initial position (x_0, y_0) to the final position (x_T, y_T) is too short for the steering angle.

In this case, this paper proposes to lengthen the distance $d_0 = \sqrt{(x_T - x_0)^2 + (y_T - y_0)^2}$ to a new distance with:

$$d_{Ni} = \rho^i d_0 \text{ with } \rho > 1 \tag{25}$$

for $i = 1, 2, 3, \ldots, n$ until the constraint in (24) is satisfied, and with ρ is an amplification coefficient, $\rho > 1$.

Then the new coordinates are:

$$x_{TN} = \rho^n (x_T - x_0), \text{ and } y_{TN} = \rho^n (y_T - y_0). \tag{26}$$

The next simulation is done with the same parameters in the previous part and subject to the constraint in Eq. (24) with an amplification coefficient $\rho = 1.1$. The constraint in (24) will be satisfied with $n = 4$ trials. The new coordinates are: $x_{TN} = y_{TN} = 23.579$.

Figure 4 shows the new trajectory and new velocity of the vehicle where the distance has been increased. The velocity of the vehicle is increased also since the final time T is not changed.

Figure 5 shows the new vehicle body angle (θ), velocity ($\dot{\theta}$) in rpm, steering angle (φ) and steering speed ($\dot{\varphi}$) in rpm for the new trajectory subject to the constraint of steer angle.

Due to the space limitations, the sideslip of this vehicle model is ignored in the present discussion. In reality, the sideslip of a vehicle depends on the tire stiffness and the cornering velocity. Then the trajectory generation of this study does not depend on the vehicle velocity. In the next part, a new vehicle trajectory generation based on polynomial distributions is investigated.

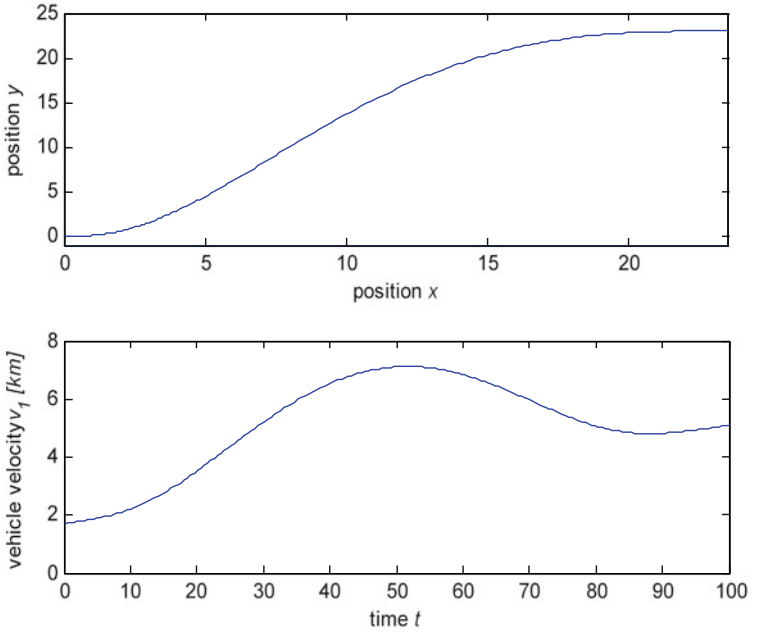


Fig. 4 Trajectory and velocity

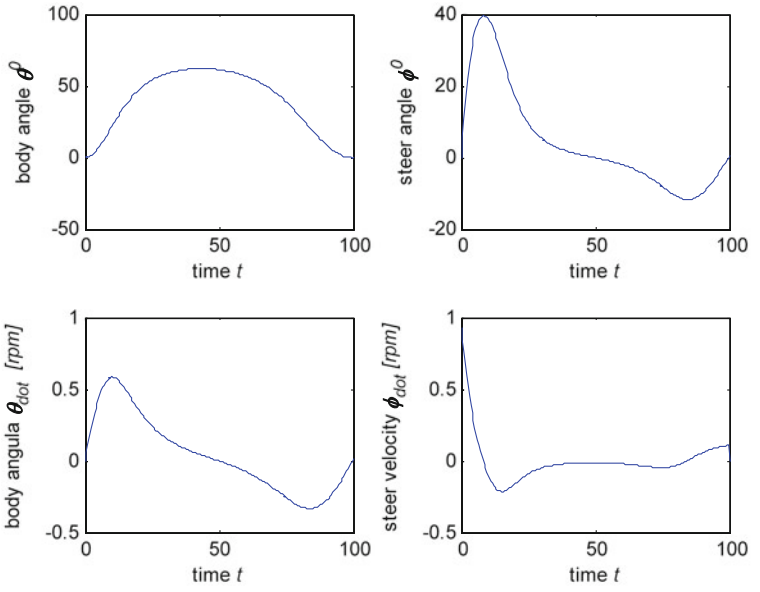


Fig. 5 Body and steer angle

5 Polynomial Trajectory Generation

For faster generation of a feasible vehicle tracking, Dong and Guo [3] have proposed a second order polynomial for flat output parameterizations. In order to do so, the Eq. (1) can be separated into the following forms:

$$z_1 = x,\ z_2 = \frac{\tan\ \varphi}{l\cos^3\ \theta},\ z_3 = \tan\ \theta \text{ and } z_4 = y \tag{27}$$

Then,

$$\dot{z}_1 = \dot{x} \tag{28}$$

$$\begin{aligned}\dot{z}_2 &= \frac{\dot{\varphi}\frac{1}{\cos^2\ \varphi} l\cos^3\ \theta + \dot{\theta} l 3\cos^2\ \theta \sin\ \theta \tan\ \varphi}{l^2\cos^6\theta} \\ &= \frac{v_2 l\cos^2\ \theta + 3rv_1\cos\ \theta\sin\ \theta\sin^2\ \varphi}{l^2\cos^5\ \theta\cos^2\ \varphi}\end{aligned} \tag{29}$$

$$\dot{z}_3 = \frac{\tan\ \varphi}{l\cos^2\ \theta} rv_1 \tag{30}$$

and

$$\dot{z}_4 = \dot{y} = \sin\ \theta rv_1 \tag{31}$$

The vehicle will move from the initial state (x_0, y_0, θ_0, φ_0) at time $t = 0$ to the final state (x_T, y_T, θ_T, φ_T) at time $t = T$ corresponding to the system in (27) from the initial state at ($z_{1.0}$, $z_{2.0}$, $z_{3.0}$, $z_{4.0}$) to the final state at ($z_{1.T}$, $z_{2.T}$, $z_{3.T}$, $z_{4.T}$).

For $0 \le t \le T$, the calculation of $[z_1(t), z_2(t), z_3(t), z_4(t)]$ will be:

$$z_1(t) = z_{1.0} + gt \tag{32}$$

$$z_2(t) = z_{2.0} + h_1 t + \frac{1}{2}h_2 t^2 + \frac{1}{3}h_3 t^3 \tag{33}$$

$$z_3(t) = z_{3.0} + gz_{2.0}t + \frac{1}{2}gh_1 t^2 + \frac{1}{6}gh_2 t^3 + \frac{1}{12}gh_3 t^4 \tag{34}$$

and

$$z_4(t) = z_{4.0} + gz_{3.0}t + \frac{1}{2}g^2 z_{2.0}t^2 + \frac{1}{6}g^2 h_1 t^3 + \frac{1}{24}g^2 h_2 t^4 + \frac{1}{60}g^2 h_3 t^5 \tag{35}$$

with, $g = \frac{z_{1.T} - z_{1.0}}{T}$, $[h_1, h_2, h_3]' = D^{-1}e$ $D = \begin{bmatrix} T & \frac{1}{2}T^2 & \frac{1}{3}T^3 \\ \frac{1}{2}gT^2 & \frac{1}{6}gT^3 & \frac{1}{12}gT^4 \\ \frac{1}{6}g^2T^3 & \frac{1}{24}g^2T^4 & \frac{1}{60}g^2T^5 \end{bmatrix}$ and

$$e = \begin{bmatrix} z_{2.T} - z_{2.0} \\ z_{3.T} - z_{3.0} - gz_{2.0}T \\ z_{4.T} - z_{4.0} - gz_{3.0}T - \frac{1}{2}g^2 z_{2.0}T^2 \end{bmatrix}.$$

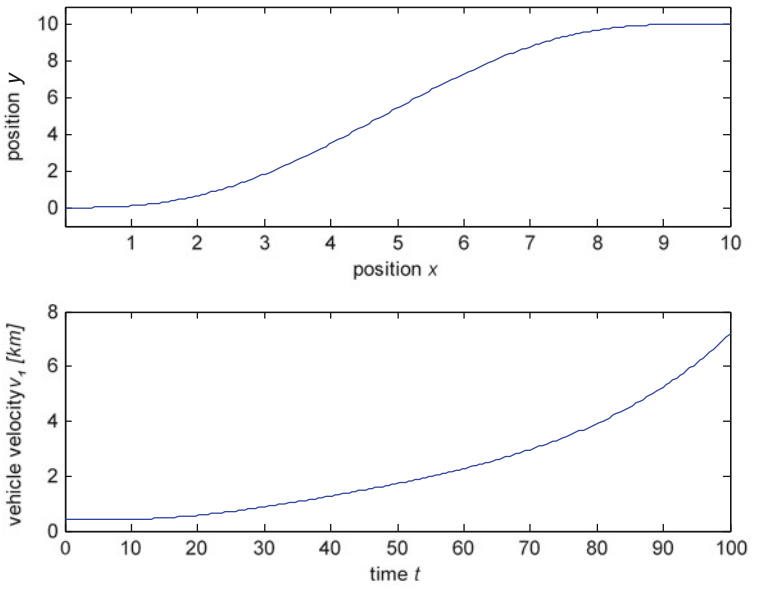


Fig. 6 Trajectory and velocity

Simulation of this trajectory generation is applied with the same parameters in Sect. 2 and shown in Figs. 6 and 7 using the following Eqs. from (36) to (42):

$$\theta = \arctan(z_3) \tag{36}$$

$$\varphi = \arctan(z_2 l \cos^3 \theta) \tag{37}$$

Putting,

$$\dot{x}(t) = \dot{z}_1(t) = g \tag{38}$$

$$\dot{y}(t) = \dot{z}_4(t) = g z_{3.0} + g^2 z_{2.0} t + \frac{1}{2} g^2 h_1 t^2 + \frac{1}{8} g^2 h_2 t^3 + \frac{1}{12} g^2 h_3 t^4 \tag{39}$$

Thus,

$$v_1(t) = \frac{\sqrt{\dot{x}^2(t) + \dot{y}^2(t)}}{r} \tag{40}$$

and

$$\dot{\theta} = \dot{z}_3 \cos^2 \theta = \left(g z_{2.0} + g h_1 t + \frac{1}{2} g h_2 t^2 + \frac{1}{3} g h_3 t^3 \right) \cos^2 \theta \tag{41}$$

and from Eq. (27)

$$\dot{\varphi} = \dot{z}_2(t) l \cos^3 \theta \cos^2 \varphi - 3\dot{\theta} \tan \theta \sin \varphi \cos \varphi \tag{42}$$

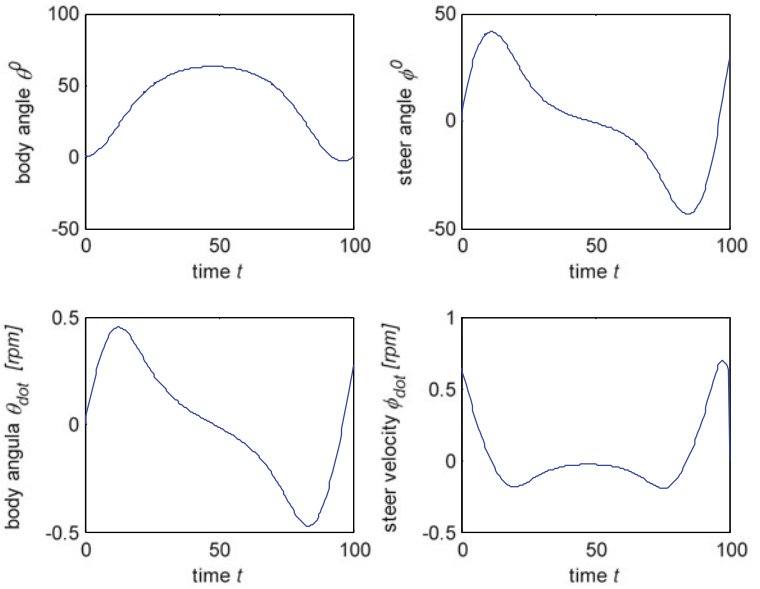


Fig. 7 Body and steer angle

Figure 6 shows the coordinate trajectory $(x,\ y)$ for the vehicle from initial point to the final point, and the velocity of the vehicle along the tracking.

Figure 7 shows the vehicle body angle,θ, and the steering angle, φ, corresponding to the angular velocity, $\dot{\theta}$, and, $\dot{\varphi}$, of the vehicle. The maximum steering angle for this trajectory generation is $\varphi = 41.5736°$ and satisfied the constraint on steering angle, $-45° \leq \varphi \leq 45°$. Therefore this trajectory generation is really better than the method presented in the previous part.From the vehicle velocity in Fig. 6, it is really not realistic to consider the vehicle speed increasing exponentially. In reality, when a vehicle is moving from one point to another point, it is better to assume that the speed will increase at the starting point and decrease at the destination point. Therefore, in the next part, a new symmetric polynomial with third order is investigated.

6 Symmetric Polynomial Trajectory Generation

Since the system is flatness and each flat output can be parameterized by a sufficiently smooth polynomials. In this part, a symmetric third order polynomial is tried for trajectory generation. Because the sideslip is ignored and then, the vehicle trajectory doesn't depend on the speed, v_1 or the travelling time, T, a new variable for this system regarding the travelling time is developed for $t = 0 \div T$:

$$x(t) = -(\frac{t}{T} - 1)^3 x_0 + \left(\frac{t}{T}\right)^3 x_T + a_x\left(\frac{t}{T}\right)^2(\frac{t}{T} - 1) + b_x\frac{t}{T}(\frac{t}{T} - 1)^2 \tag{43}$$

and the trajectory of y is:

$$y(t) = -(\frac{t}{T} - 1)^3 y_0 + \left(\frac{t}{T}\right)^3 x_T + a_y\left(\frac{t}{T}\right)^2(\frac{t}{T} - 1) + b_y\frac{t}{T}(\frac{t}{T} - 1)^2 \tag{44}$$

derivation of $x(t)$:

$$\dot{x}(t) = -3(\frac{t}{T} - 1)^2 x_0 + 3\left(\frac{t}{T}\right)^2 x_T + a_x 2\frac{t}{T}(\frac{t}{T} - 1) + a_x\left(\frac{t}{T}\right)^2 + b_x(\frac{t}{T} - 1)^2 + b_x 2\frac{t}{T}(\frac{t}{T} - 1) \tag{45}$$

and derivation of $y(t)$:

$$\dot{y}(t) = -3(\frac{t}{T} - 1)^2 y_0 + 3\left(\frac{t}{T}\right)^2 y_T + a_y 2\frac{t}{T}(\frac{t}{T} - 1) + a_y\left(\frac{t}{T}\right)^2 + b_y(\frac{t}{T} - 1)^2 + b_y 2\frac{t}{T}(\frac{t}{T} - 1) \tag{46}$$

Then,

$$\ddot{x}(t) = -6(\frac{t}{T} - 1)x_0 + 6\frac{t}{T}x_T + a_x 2(2\frac{t}{T} - 1) + a_x 2\frac{t}{T} + b_x 2(\frac{t}{T} - 1) + b_x 2(2\frac{t}{T} - 1) \tag{47}$$

and,

$$\ddot{y}(t) = -6(\frac{t}{T} - 1)y_0 + 6\frac{t}{T}y_T + a_y 2(2\frac{t}{T} - 1) + a_y 2\frac{t}{T} + b_y 2(\frac{t}{T} - 1) + b_y 2(2\frac{t}{T} - 1) \tag{48}$$

The constraint on speed:

$$rv_1 = \frac{\dot{x}}{\cos\,\theta} = \frac{\dot{y}}{\sin\,\theta} \tag{49}$$

The constraint at starting point $t = 0$:

$$\dot{x}(0) = k_0 \cos\,\theta_0, \text{ and } \dot{y}(0) = k_0 \sin\,\theta_0 \tag{50}$$

The constraint at destination point $t = T$:

$$\dot{x}(T) = k_T \cos\,\theta_T, \text{ and } \dot{y}(T) = k_T \sin\,\theta_T \tag{51}$$

From Eqs. (46), (47), (49) and (50), for the calculation simplicity, it is assumed that the speed coefficients at the start and destination point, $k_0 = k_T = k$, then

$$a_x = k\cos\,\theta_T - 3x_T, \text{ and } b_x = k\cos\,\theta_0 - 3x_0 \tag{52}$$

Similarly,

$$a_y = k \sin\ \theta_T - 3y_T, \text{ and } b_y = k \sin\ \theta_0 - 3y_0 \tag{53}$$

Simulation of this symmetric polynomial is conducted with the same parameters in the previous parts and shown in Figs. 8 and 9. The speed coefficients are $k_0 = k_T = k = 1$. Other parameters regarding the vehicle velocity, $v_1(t)$; the body angle, $\theta(t)$, and the body angle velocity, $\dot{\theta}(t)$; the steering angle, $\varphi(t)$, and the steering angle velocity, $\dot{\varphi}(t)$ are calculated in the following equations:

$$v_1 = \frac{\sqrt{\dot{x}^2 + \dot{y}^2}}{r} \tag{54}$$

$$\theta = \arctan\left(\frac{\dot{y}}{\dot{x}}\right) \tag{55}$$

$$\varphi = \arctan\left(l\frac{\cos^3\theta\ddot{y}}{(\dot{x})^2}\right) \text{ or } \varphi = \arctan\left(l\frac{\ddot{y}\dot{x} - \ddot{x}\dot{y}}{r(\dot{x}^2 + \dot{y}^2)^{\frac{3}{2}}}\right) \tag{56}$$

$$\dot{\theta} = \frac{\ddot{y}\dot{x} - \ddot{x}\dot{y}}{\dot{x}^2}\frac{1}{\left(\frac{\dot{y}}{\dot{x}}\right)^2 + 1} = \frac{\ddot{y}\dot{x} - \ddot{x}\dot{y}}{\dot{x}^2 + \dot{y}^2} = \frac{\tan\ \varphi}{l} r v_1 \tag{57}$$

and

$$\dot{\varphi} = \frac{\partial\left(\arctan\left(l\frac{\ddot{y}\dot{x} - \ddot{x}\dot{y}}{r(\dot{x}^2 + \dot{y}^2)^{\frac{3}{2}}}\right)\right)}{\partial t} \tag{58}$$

It can be seen from Fig. 8 that the trajectory of this symmetric polynomial is more realistic because the vehicle velocity increases from the starting point and decreases in the destination point.

As shown in Fig. 9, the maximum steering angle for this third order symmetric polynomial method is $\varphi = 41.1622°$ and satisfied the constraint on steering angle as $-45° \leq \varphi \leq 45°$. The steer angle is still a little bit smaller than the steer angle in the previous part for the second order polynomial method, $\varphi = 41.5736°$.

In the next part, an analysis and comparison of the above three methods are presented with simulations for the vehicles moving in both forward and backward directions.

7 Performance Analysis

Performance comparisons of the three methods on the trajectory generation and the vehicle velocity are shown in Figs. 10 and 11.

From Fig. 10, it can be seen that the polynomial method can produce a trajectory with smoother path. A symmetric polynomial generation can produce a

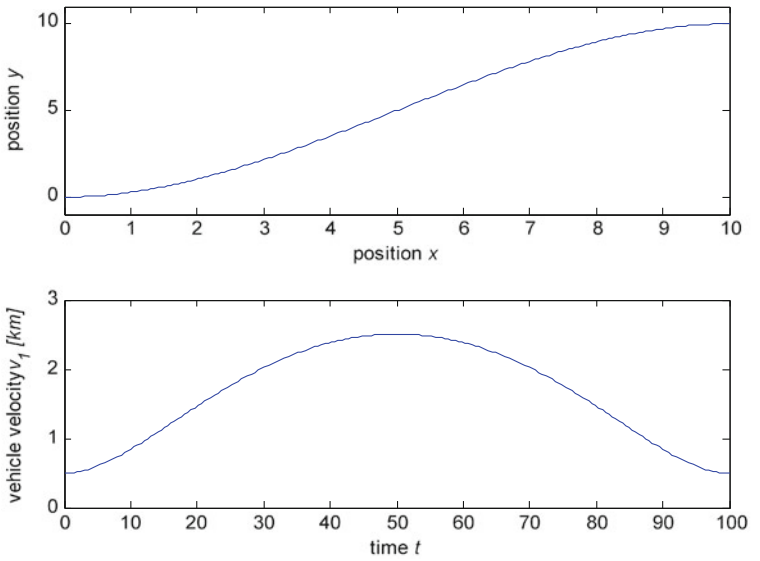


Fig. 8 Trajectory and velocity

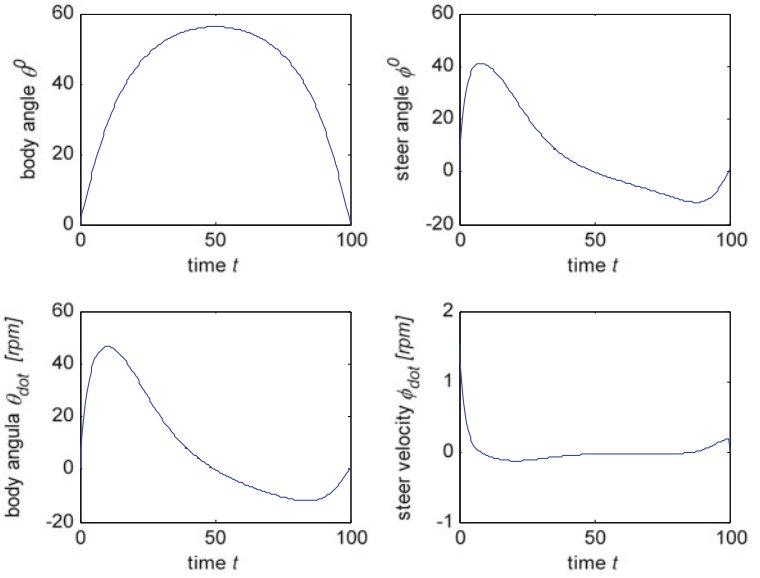


Fig. 9 Body and steer angle

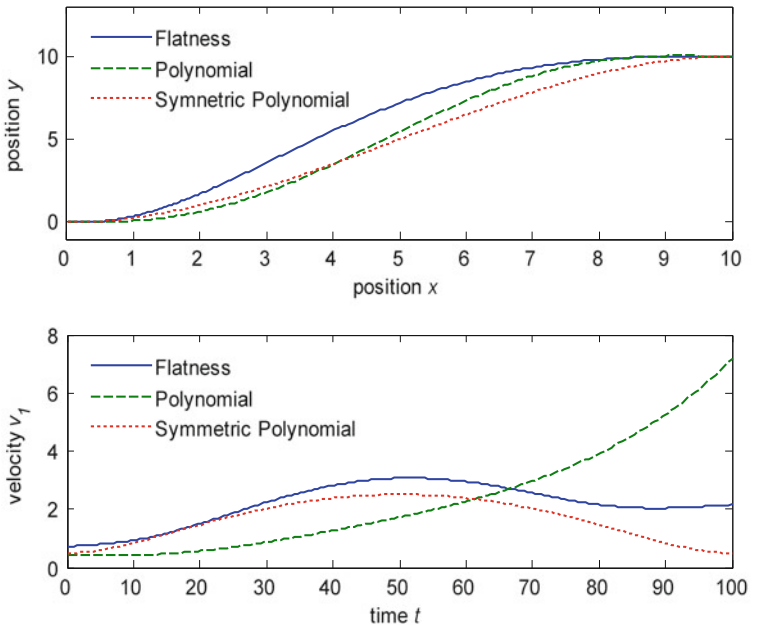


Fig. 10 Comparison of trajectory and velocity

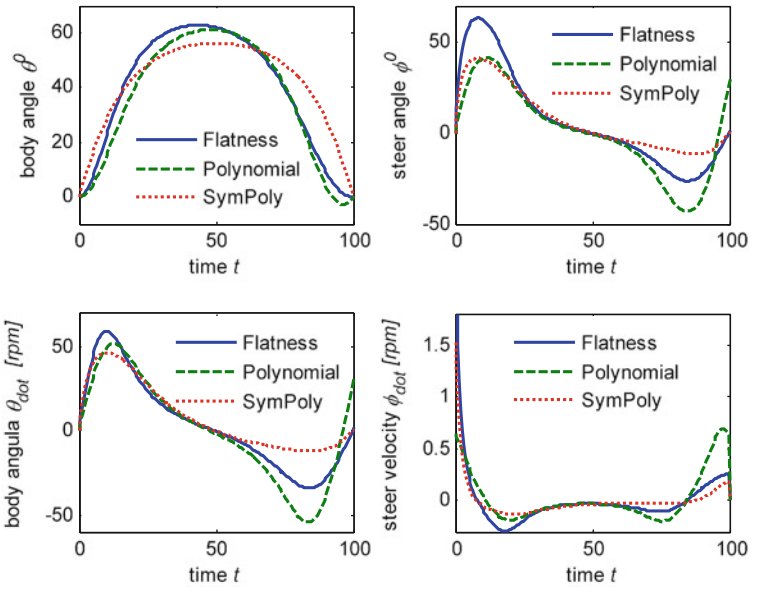


Fig. 11 Comparison of body and steer angle

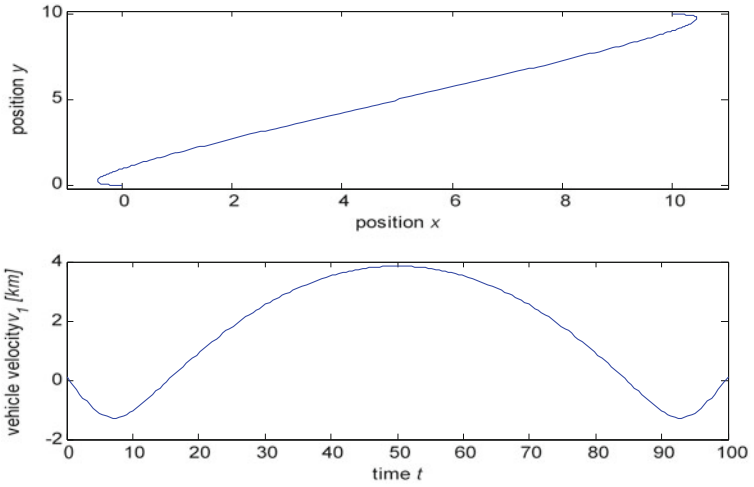


Fig. 12 Trajectory and velocity

more realistic speed since it allows the vehicle gradually increases the speed at the start point and reduces the speed at the destination point.

Figure 11 shows that the third order symmetric polynomial method can provide the lowest body angle, $\theta(t)$, the body angle velocity, $\dot{\theta}(t)$; the steering angle, $\varphi(t)$, and the steering angle velocity, $\dot{\varphi}(t)$. Then, this method is recommended for the development of an automatic control of tracking vehicles.

Studies for the vehicle moving in reverse speeds are also conducted. As mentioned in Eq. (3), the vehicle velocity, v_1, in reverse speeds will change the sign. Then the speed coefficients in Eqs. (52) and (53) will be minus values. Simulations for the vehicle in reverse speeds are done with the above third order symmetric polynomial trajectory generation. The speed coefficients now are assigned as, $k_0 = k_T = k = -1$. Results of the simulation are shown in Figs. 12 and 13.

Since the vehicle is forced to reverse at the start point, for $k_0 = -1$, and at the destination point, for $k_T - 1$, Fig. 12 shows the vehicle backing out at the start point, going forward to the destination, then backing to the destination parking space. The vehicle velocities are changing the direction in three times.

Figure 13 shows the body angle, θ, switching 180° in two times corresponding to each reverse speeds. The maximum steering angle in this reverse trajectory is, $\varphi_{\max} = 33.7097°$, and much lower than the required angles in forward movements. The body angular velocity, $\dot{\theta}$, as well as the steering angular velocity, $\dot{\varphi}$, are also indicated during this reverse performance.

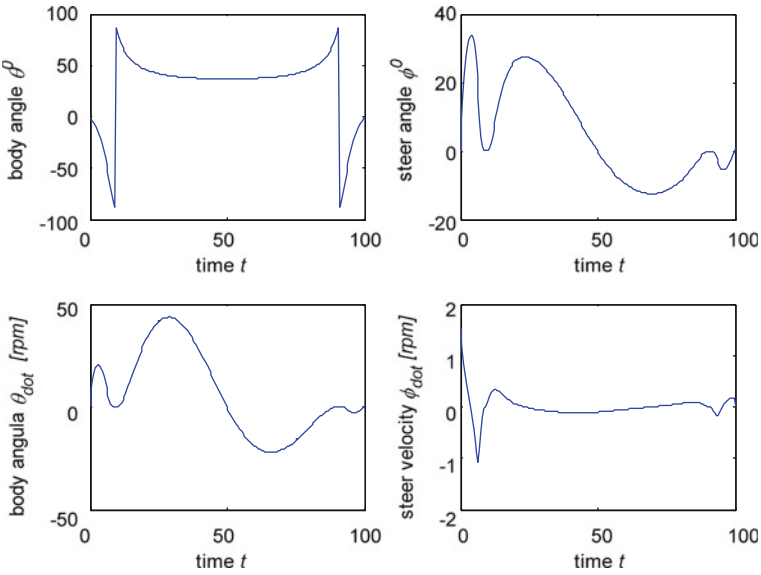


Fig. 13 Body and steer angle

8 Conclusion

The chapter has presented three methods of trajectory generation for autonomous vehicles subject to constraints. Regarding to the real vehicle speed development, the third order symmetric polynomial trajectories are recommended. Simulations and analyses are also conducted for vehicle with forward and reverse speeds. Results from this study can help to develop a real-time control system for auto-driving and auto-parking vehicles. The limitation of this study is to ignore the influence of the vehicle sideslip due to the cornering velocity. However this error can be eliminated with the feedback control loop and the offset margin of the steer angle allowance.

References

1. J. Choi, R. Curry, G. Elkaim, Curvature-Continous trajectory generation with corridor constraint for automomous ground vehicles, in *Proceedings of IEEE Conference of Decision and Control*, Atlanta, Georgia, 15–17 Dec 2010
2. V. Delsart, T. Fraichard, L. Martiez, Real-Time trajectory generation for car-like vehicles navigating dynamic environments,in *proceedings of IEEE International Conference on Robotics and Automation*, Kobe, Japan, 12–17 May 2009
3. W. Dong, Y. Guo, New trajectory generation methods for nonholonomic mobile robots, in *Proceedings of the International Symposium on Collaborative Technologies and Systems*, Missouri, USA, 20–20 May 2005

4. M. Gomez, Optimal control for wheeled mobile vehicles based on cell mapping techniques, in *Proceedings of IEEE Intelligent Vehicles Symposium,* Eindhoven, Netherland, 4–6 June 2008
5. K. Kanjanawanishkul, M. Hofmeister, A. Zell, Smooth reference tracking of a mobile robot using nonlinear model predictive control, in *Proceedings of the 4th European Conference on Mobile Robots,* Mlini/Dubrovnik, Croatia, 23–25 Sept 2009
6. G. Klancar, Tracking-error model-based predictive control for mobile robots in real time. Robot. Auton. Syst. **55**(6), 460–469 (2007)
7. J. Levine, *Analysis and control of nonlinear system, a flatness-based approach* (Springer, Heidelberg, 2009)
8. VT. Minh, Vehicle steering dynamic calculation and simulation, in *Proceedings of the 23rd International DAAAM Symposium on Intelligent Manufacturing and Automation*, Zadar, Croatia, 24–27 Oct 2012
9. V.T. Minh, Stabillity for switched dynamic hybrid systems. Math. Comput. Modell. **57**(1–2), 78–83 (2013)
10. V.T. Minh, H. Fakhruldin, Time forward observer based adaptive controller for a teleoperation system. Int. J. Control Autom. Syst. **9**(3), 470–477 (2011)
11. V.T. Minh, A. Nitin, A comparative study on computational schemes for nonlinear model predictive control. Asian. J. Control. **8**(4), 324–331 (2006)
12. M. Missura, S. Behnke, Efficient kinodynamic trajectory generation for wheeled robots, in *Proceedings of IEEE International Conference on Robotics and Automation*, Shanghai, China, 9–13 May 2011
13. J. Wang, Z. Lu, W. Chen et al, An adaptive trajectory tracking control of wheeled mobile robots, in *Proceedings of IEEE Conference on Industrial Electronics and Applications*, Beijing, China, 21–23 June 2011
14. M. Werling, S. Kammel, J. Ziegler et al., Optimal trajectories for time-critical street scenarios using discretized terminal manifolds. Int. J. Robot. Res. **31**(3), 346–359 (2011)
15. T. Zhou, M. Li, Z. Mai et al, Trajectory generation model-based IMM tracking for safe driving in intersection scenario. Int. J. Veh. Technol.(2011). doi:10.1155/2011/103696

Fuzzy Logic-Based Adaptive Decision Support in Autonomous Vehicular Networks

Chrysostomos Chrysostomou, Constantinos Djouvas and Lambros Lambrinos

Abstract The area of intelligent autonomous vehicles and systems poses new challenges in providing mechanisms for efficient communication and control between vehicles, as well as developing robust, adaptive techniques to support intelligent transportation system applications. In this chapter, we show the need for providing an intelligent controller offering decision support in autonomous vehicular networks in terms of broadcast communication channel access. Specifically, we exploit fuzzy logic control, derived from its reported strength of using linguistic information to control nonlinear systems, to build an adaptive, intelligent controller, based on the traffic density, to aid vehicles in deciding when to access the broadcast communication channel. It is demonstrated, by means of enriched simulative evaluation that the fuzzy logic-based controller offers inbuilt robustness with effective control of the system under dense conditions, in contrast with the conventional—IEEE 802.11p standard—solution we compared against.

Keywords Autonomous vehicular networks · Fuzzy logic control · Decision support · Cooperative awareness · Broadcast communication channel access

C. Chrysostomou (✉)
Department of Computer Science and Engineering, Frederick University, Nicosia, Cyprus
e-mail: ch.chrysostomou@frederick.ac.cy

C. Djouvas · L. Lambrinos
Department of Communication and Internet Studies, Cyprus University of Technology, Limassol, Cyprus
e-mail: costas.tziouvas@cut.ac.cy

L. Lambrinos
e-mail: lambros.lambrinos@cut.ac.cy

Z. H. Khan et al. (eds.), *Computational Intelligence for Decision Support in Cyber-Physical Systems*, Studies in Computational Intelligence 540,
DOI: 10.1007/978-981-4585-36-1_7,

1 Introduction

Intelligent vehicular networks represent a challenging area that enables vehicles to communicate with each other (vehicle-to-vehicle communication) and/or with roadside infrastructure (vehicle-to-infrastructure communication). The aim is the facilitation of a plethora of applications, such as traffic safety, traffic efficiency and management, as well as infotainment. Particularly, vehicular networks support Intelligent Transportation System (ITS) applications, which aim at providing safety, efficiency, and comfort of future road traffic.

In the case of autonomous vehicles comprising a vehicular network, the aforementioned aims are crucial. Cooperative awareness is essential among autonomous vehicles, from which many applications can be derived, like reducing traffic jams and avoiding vehicle collisions. To build cooperative awareness, each vehicle periodically sends a small status message, known as a beacon message. For this, processing and communication facilities are located in a vehicle, known as onboard units, providing an application runtime environment, positioning, security and communication functions and interfaces with other vehicles [1].

For communication in vehicular networks, the Wireless Access for the Vehicular Environment (WAVE) standards have been proposed [2–5]. The IEEE 802.11 standard body has proceeded to a new amendment, notably, the IEEE 802.11p [6, 7]. The allocation of the Dedicated Short Range Communications (DSRC) spectrum band [8, 9] to be used exclusively for vehicular environments was followed by the IEEE 802.11p WAVE standardization process [10].

The IEEE 802.11p medium/channel access control (MAC) protocol is based on the carrier sense multiple access with collision avoidance (CSMA/CA) MAC protocol. With the aid of the MAC protocol, a vehicle can decide when to access the communication channel based, among other parameters, on the contention window (CW) size that is doubled when a collision occurs, in the case of unicast communication. Moreover, the IEEE 802.11p employs the enhanced distributed channel access (EDCA) mechanism, which is originally provided by the IEEE 802.11e [11], to differentiate traffic types through different static MAC parameters values.

However, due to the dynamics of the vehicular environment, vehicular density may vary greatly and this may result in traffic density variation. Thus, the communication channel may become congested. It has been observed [12, 13] that as the traffic density increases, the load on the channel can increase radically. This will have as a consequence, the deterioration of the performance of the ITS applications.

In the literature, several research works evaluate the IEEE 802.11p MAC protocol performance in unicast communication and enhancements are proposed [14–18]. However, in broadcast communication, as is the case of the active road safety and control applications along with the use of beacons for cooperative awareness among autonomous vehicles, there is further a need to improve the performance of the MAC protocol. This is due to the fact that because of the lack

of acknowledgements when performing broadcast beaconing and other active road safety/control messages, the CW MAC parameter is never increased. Thus, despite a possible increase of the traffic load and consequently the severe congestion on the communication channel, the MAC protocol does not respond appropriately. This behavior badly influences the communication between the vehicles, especially when safety applications are run, like vehicle collision avoidance etc.

Consequently, an alternative to the conventional solution for controlling the broadcast communication channel is the employment of fuzzy logic control (a well-known Computational Intelligence technique) due to its inherent robustness and ability to cope with the uncertainties and non-linearities of the system to be controlled. We propose that fuzzy logic control should have an essential role to play in designing this challenging broadcast decision support/control system.

The chapter is organized as follows. Section 2 gives a brief outline of the wireless access in vehicular environments and the 802.11p protocol. In Sect. 3 the related work is presented, and in Sect. 4 we briefly introduce the characteristics of fuzzy logic control and its application in network control problems. In Sect. 5, we describe how the proposed fuzzy logic based controller for adaptive decision support can be applied in vehicular networks. Then, in Sect. 6 the simulative performance evaluation of the proposed mechanism compared with the IEEE 802.11p standard protocol is discussed. Finally, in Sect. 7, we present our conclusions.

2 Wireless Access in Vehicular Environments

Vehicular networks characterize a principally demanding class of mobile (ad hoc) networks that enable vehicles to communicate with each other (vehicle-to-vehicle communication) and/or with roadside infrastructure (vehicle-to-infrastructure communication). Active road safety, traffic efficiency and management, as well as infotainment are prime classes of the plethora of applications whose development and deployment is based upon the exploitation of vehicular networks.

Specifically, active road safety applications are mainly employed to decrease the probability of accidents and the loss of life of the occupants of vehicles [1]. Examples of such applications are [1]: intersection collision warning, lane change assistance, overtaking vehicle warning, cooperative forward collision warning, pre-crash sensing/warning, and hazardous location notification. Further to these, traffic efficiency and management applications focus on improving the vehicle traffic flow, traffic coordination and traffic assistance, like speed management and cooperative navigation [1]. Also, infotainment applications include cooperative local services, as well as global Internet services, such as media downloading and financial services [1].

Due to the diversity in requirements between these types of applications, the perceived Quality of Service (QoS) varies; a timing failure might have no effect (non-real-time), might compromise service quality (soft real-time) or might lead to

a disaster (hard real-time). For this, the IEEE has developed the WAVE system architecture [2–5] to provide wireless access in vehicular environments; the IEEE 802.11p [6] and IEEE 1609.x [2–5] standards are the building blocks of this architecture as their goal, as a whole, is to aid the provision of wireless access in vehicular environments [10].

A WAVE system consists of two main entities:

- Roadside units (RSUs) include equipment located along highways, at traffic intersections and at other locations where timely communication with vehicles is needed [1].
- Onboard units (OBUs) are processing and communication facilities located inside a vehicle, providing an application runtime environment, positioning, security and communication functions and interfaces with other vehicles [1].

WAVE units operate independently exchanging information over the control channel (CCH) which is a fixed radio channel. Additionally, they may also organize themselves in small networks called WAVE basic service sets (WBSSs) which can consist of OBUs only or a mix of OBUs and RSUs. All the members of a particular WBSS (the provider that initiates the communication and the possible users) exchange information via one of a number of radio channels known as service channels (SCHs) [10]. Generally, the CCH is reserved for system control and safety messages, whereas the SCHs are used to exchange non-safety data.

The IEEE 802.11p is based on the carrier sense multiple access with collision avoidance (CSMA/CA) MAC protocol. It employs the enhanced distributed channel access (EDCA) mechanism, which is originally provided by the IEEE 802.11e [11], to distinguish between different traffic types through different values for the static MAC parameters (e.g. traffic prioritization through different CW_{min} and CW_{max} values for different traffic types).

Four applications' access categories (ACs) are defined in the WAVE standards. The differentiation in priority between ACs for channel access parameters is implemented using the appropriate EDCA parameter set values (see Table 1 that refers to the use of the CCH by active road safety and control applications), which are defined as follows [5]:

- Arbitration inter-frame space (AIFS): the minimum time interval between the wireless medium becoming idle and the start of the next frame transmission.
- Contention window (CW): An interval out of which a random number is drawn to implement the random back-off mechanism.
- aCW_{min} and aCW_{max}, which are static values (15 and 1023, respectively), as specified in [6].

When an emergency event occurs, the communication channel utilization is likely to degrade due to substantial broadcast of emergency messages. Therefore, there is a need to prioritize messages based on the application requirements. Data packets carrying traffic safety and traffic efficiency information usually have higher significance, and therefore should have access to the communication channel

Table 1 Default EDCA parameter set used on the CCH [5]

AC index	AC	CW_{min}	CW_{max}	AIFSN
1	Background	aCW_{min}	aCW_{max}	9
0	Best effort	$(aCW_{min} + 1)/2-1$	aCW_{min}	6
2	Video	$(aCW_{min} + 1)/4-1$	$(aCW_{min} + 1)/2-1$	3
3	Voice	$(aCW_{min} + 1)/4-1$	$(aCW_{min} + 1)/2-1$	2

quicker (i.e. have higher priority) than background or control messages such as periodic beacon messages [1]. With the employment of the IEEE 802.11p MAC protocol, this differentiation/prioritization is defined (as seen from Table 1).

Note that the back-off mechanism is used in case the communication channel is sensed busy, and a back-off time is chosen uniformly at random from the interval [0, CW + 1], where the initial CW is equal to the CW_{min}. In the case of unicast communication, the interval size gets doubled, until it's equal to CW_{max}, if the subsequent transmission attempt fails/collides (i.e. no acknowledgment is received). However, in the case of broadcast communication, the CW parameter is never increased due to the lack of acknowledgements when performing broadcast beaconing and other active road safety/control messages. Due to the dynamics of the vehicular environment, vehicular density may vary greatly and hence traffic density too; the communication channel may become congested as the load on the channel increases radically, and consequently the performance of the ITS applications may deteriorate. Thus, despite a possible increase of the traffic load and consequently the severe congestion on the communication channel, the MAC protocol does not respond appropriately. This behavior badly influences the communication between the vehicles, especially when safety applications are run, like vehicle collision avoidance etc.

3 Related Work

In the literature, several research works evaluate the IEEE 802.11p MAC protocol performance in unicast communication with regard to the access of the communication channel, and enhancements are proposed (e.g. [14–18]). Specifically the impact of using/adapting different CW settings on the performance of the MAC protocol has been examined.

In [14], the outcome of the evaluation of the 802.11p protocol is that the AC for high priority traffic suffers from a degradation of the performance as the traffic density increases. However, this study only evaluates a single AC at a time.

In [15] two schemes are presented that adapt the CW, based on the number of transmitting vehicles. The first proposal is a centralized approach; it assumes prior knowledge on the number of transmitting vehicles something which is difficult to obtain due to the variability and high dynamics of the environment. The second proposal is a distributed approach: vehicles use local channel information to adapt

the CW size. A linear updating method is used to change the back-off window size based on channel busy measurements; however, the high dynamics of the environment imply that such a linear updating method is not robust enough. The evaluation of these schemes has taken into consideration only one application AC for each scenario, ignoring the mixture of prioritized traffic types expected to be present.

A decentralized self-organizing time division multiple access (STDMA) scheme is presented in [16]. Although this scheme has its merits, the arguably limited capabilities of TDMA MAC schemes will be a serious impediment in deploying a wide range of applications in vehicular networks. Furthermore, the provision of effective differentiation is not examined as in the evaluation of the proposed scheme all data packets have the same priority.

In [17], a fuzzy logic based enhanced 802.11p mechanism is proposed which applies a non-linear control law to adapt the back-off/CW parameter for each transmitting node; using the channel traffic occupancy, in combination with the applications' ACs, it aims to be adaptive to the high environment dynamics. Further, in [18], an adaptive, fuzzy-logic based, medium access control mechanism is shown, in which the minimum and maximum CW parameters are dynamically tuned based on network measurements, concerning all applications' ACs. Through simulative evaluation, both techniques have achieved adequate QoS provision in terms of throughput performance, by providing effective differentiation among different applications' ACs.

All these techniques are implemented only in cases of unicast communication. However, in broadcast communication, as is the case of the active road safety and control applications along with the use of beacons for cooperative awareness among autonomous vehicles, there is further a need to improve the performance of the underlying MAC protocol. This is due to the fact that because of the lack of acknowledgements when performing messages broadcasting, the CW MAC parameter is never increased. Thus, despite a possible increase of the traffic load and consequently the severe congestion on the communication channel, the MAC protocol does not respond appropriately. This has a negative impact on the communication between the vehicles, especially when safety applications are run.

In [13], the authors provide an analysis of the effects of the CW size on the performance of broadcast beaconing in vehicular networks. They have shown that increasing the CW size as the traffic density increases does not improve beaconing performance. However, it should be noted that in that paper all nodes in a scenario used the same CW settings. Also, no mobility of the nodes is provided.

Consequently, encouraged by the solid work proposed in [17, 18], an alternative to the conventional solution for controlling the broadcast communication channel is the employment of fuzzy logic control (a well-known Computational Intelligence technique) due to its ability to cope with the uncertainties and non-linearities of the system to be controlled and its inherent robustness that provides. We propose that fuzzy logic control should have an essential role to play in designing this challenging broadcast decision support/control system.

4 Fuzzy Logic Control: Theory and Applications

4.1 Fuzzy Logic Control

Fuzzy logic [19] is one of the Computational Intelligence tools, which is used to create a multivalued logic system. It is based on fuzzy set theory [19], where a fuzzy set is described with the aid of a membership function, which can be any real number in the normalized interval [0, 1]. This expresses the grade of membership an element can belong to a specific fuzzy set.

Fuzzy Logic Control (FLC) [20] refers to the use of fuzzy sets together with the employment of fuzzy inference in order to develop control laws in the form of "IF–THEN" linguistic rules, by utilizing the understanding of expert(s) in the field of the application of concern. This is of particular interest, as Fuzzy logic employs expert(s) and/or operator's qualitative experience to build the control law. The defined set of linguistic rules describes what action should be taken based on observations of the behavior of the controlled system.

FLC has been successfully established [21, 22] as the means of designing controlled systems, especially in the cases where it is hard to mathematically model such systems. Thus, the difficulty in designing a formal analytical model when techniques from control theory are too complicated to be used and/or the model itself is highly nonlinear, allows the design of feedback controllers, when an intuitive understanding of the system under control is available, with the aid of FLC. There is a lot of attention on the use of FLC by the research community, both academia [23, 24] and industry [25]. This leads to a successful exploitation of FLC to qualitatively capture the characteristics of a system, based on intuitive understanding and system behavior observations, in order to offer control in complex systems (e.g. [26]).

Therefore, a fuzzy logic controller can be considered as a nonlinear controller that is illustrated with the aid of linguistic rules that describe the relationship between the input(s) and output(s) of the controller. With the use of FLC the control algorithm can be easily modified/tuned and better realized. Moreover, a fuzzy logic controller is independent of any mathematical modeling, thus it diminishes any complexity in the controller's design, and further, it attains inbuilt robustness.

4.2 Network Control Applications

The exploitation of FLC in a wide number of fields is evident by the research papers found in the literature that utilize fuzzy logic to provide solutions in a variety of problems. Specifically, there are numerous published research papers that deal with the provision of congestion control—decision support in computer networking. For example, FLC has been used in solving congestion control

problems in Asynchronous Transfer Mode (ATM) networks (e.g. [23, 24, 27–29]), since it was difficult to design formal analytical models for such complex systems.

FLC was also used in the Internet Protocol (IP) networks, as is the case of the successful deployment of FLC in the active queue management (AQM) problem in IP networks [30–33]. A comprehensive study can be found in [34].

As stated above, due to the enormous experience of successful employment of FLC in designing effective control laws that control complex and highly nonlinear systems, the deployment of fuzzy control techniques to the problem of congestion control—decision support in computer networks is worthy of further investigation. Recently, fuzzy logic has been successfully used in controlling the access to the communication channel, in terms of unicast communication, in vehicular networks [17, 18] (already discussed in Sect. 3).

5 Fuzzy Logic-Based Adaptive Decision Support

In this chapter a Fuzzy Logic-based controller for Adaptive Decision Support (FLADS) regarding the access of the broadcast communication channel in vehicular networks is presented. The vehicles need to exchange beacon messages as part of the cooperation awareness to fulfill the requirements of ITS applications, like reducing traffic jams, vehicle collision avoidance, along with the broadcast of any active road safety and control messages. The aim of the proposed controller is to overcome the problems the conventional channel access control mechanism (i.e. the IEEE 802.11p protocol) has; that is, it does not behave adequately, especially in broadcast communication (as discussed in Sect. 3). A simple and efficient controller is designed, utilizing FLC principles; thus, a nonlinear control law is constructed to provide effective control of the system with inbuilt robustness. Although the main focus is on the development of a broadcast channel access control scheme for decision support in vehicular networks to support active road safety and traffic efficiency and management applications that reserve the CCH, the same principles can be easily adopted for unicast communication in vehicular networks.

It is a necessity for the adaptive channel access control mechanism to operate in a decentralized and self-organized way, i.e. locally at each autonomous vehicle. This is due to the fact that the network environment is highly dynamic with high mobility. The proposed scheme adapts the back-off/contention window parameter for each transmitting vehicle based on the channel traffic occupancy, combined with the access categories the applications belong to, in order to be adaptive to the high environment dynamics; thus, the appropriate decision support is provided to the autonomous vehicles to control the wireless access using linguistic rules that describe the behavior of the environment in differing widely operating conditions. The perspective achievement is the Quality of Service provision in terms of throughput performance, as well as delay and message losses. Thus, the proposed mechanism alleviates the problem encountered when the conventional solution is

used, i.e. the CW parameter never increases despite the severity of the communication broadcast channel conditions.

In order to design the FLADS controller the following standard steps were followed [34]:

- Identification of the inputs and output and their ranges.
- Construction of the rule base (knowledge) that the system will operate under.
- Design of the degree of each input and output fuzzy membership function.
- Design of the inference engine.
- Selection of the proper defuzzification method on the output.

In order to design the fuzzy logic-based controller, there is no regular procedure. In this chapter, we adopted the successful, simple, and most commonly used procedure that was utilized in our earlier work [34]. We first need to qualitatively understand the system to be controlled, and then we define the membership functions of the inputs and output, as well as we construct the rule base. If needed, a trial and error approach is followed in order to test/tune the system, until we accomplish satisfactory performance, and at the same time we keep the controller as simple and generic as possible. It is worth noting that it is very hard to study analytically the influence of certain parameters of the controller, since the constructed fuzzy logic based controller is nonlinear. The past experience indicates that extensive simulation and experimentation are enough to be used to examine the behavior of the controller.

The important steps followed to design the FLADS controller are briefly described below.

5.1 Decision on the Input and Output

As indicated in [20, 34], the use of multiple inputs can enhance the ability to capture the dynamic state of the system to be controlled, as well as to improve the linguistic description of the dynamics of the system. For the construction of the FLADS controller, we adopt the simplest Multiple Input Single Output (MISO) model, by designing a two-input, single-output fuzzy logic based controller on each autonomous vehicle in vehicular networks.

It is of high importance to select the proper inputs and output of the FLADS controller that can be adopted in any network/traffic condition. Thus, we have chosen to utilize the channel traffic occupancy measured for two successive sampling intervals (this can be interpreted as a prediction horizon). Based on the channel condition the appropriate decision support is dynamically offered in terms of the access of the broadcast communication channel. Also, the output of the controller is selected as a nonlinear back-off/contention window parameter that is used as the input to the controlled system in order to adaptively decide when the communication channel can be accessed. Further, the proposed controller supports

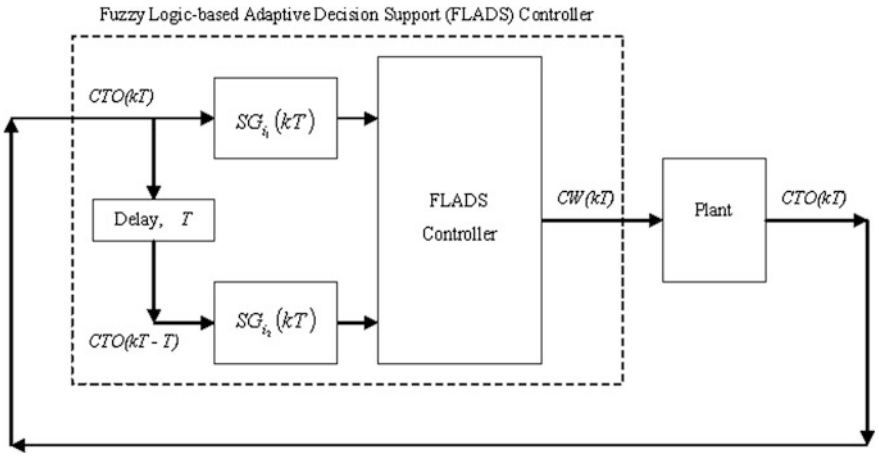


Fig. 1 Fuzzy logic-based adaptive decision support (FLADS) system model

the prioritization offered to the applications, based on the defined applications' ACs (as discussed in Sect. 2—Table 1). Thus, for each AC that has different minimum and maximum CW parameter values, the appropriate FLADS controller runs.

Based on the above decisions, the fuzzy control system, for the proposed FLADS controller, is specified (see Fig. 1), where all quantities are considered at the discrete instant kT:

- T is the sampling period.
- *CTO(kT)* is the channel traffic occupancy, measured throughout the current sampling period, by keeping the amount of time a channel is busy. The record of busy time within each sampling period is achieved by physical layer channel idle/busy indication.
- *CTO(kT-T)* is the channel traffic occupancy, measured at the previous sampling period.
- *CW(kT)* is the calculated contention window parameter used in determining the back-off time.
- $SG_{i1,2}(kT)$ are the input scaling gains.

As shown in Fig. 1, the FLADS dynamically calculates the contention window parameter based on a nonlinear control law derived by the construction of the fuzzy logic system, and taking into account the density of the environment. This is in contrast with the basic operation of the IEEE 802.11p in broadcast communication, where the contention window is kept at its minimum value, without taking into account the density of the medium. Further, due to the high variability and dynamics of the system, a nonlinear control law operates more efficiently in comparison with a linear control method.

The range of the input and output values for a given controller is usually called the "universe of discourse" [20]. Usually, in order to provide a generic fuzzy logic

based controller, the universe of discourse for each input is normalized to the interval [0, 1] by means of constant scaling factors [20]. For the FLADS controller design, the input scaling gains, $SG_{i1,2}(kT)$, shown in Fig. 1, are selected so as the range of values of $SG_{i1}(kT)$* CTO(kT), and $SG_{i2}(kT)$* CTO(kT-T) lie within [0, 1]. Based on the above, the input scaling gains $SG_{i1,2}(kT)$ are set as shown in Eq. (1):

$$SG_{i1,2}(kT) = 1/T \tag{1}$$

The range of values of the controller's output, *CW(kT)*, lies between aCW_{min} and aCW_{max}, according to the access category the application, running on each autonomous node, belongs to (see Sect. 2—Table 1).

5.2 Decision on the Rule Base

Linguistic rules are used to dynamically calculate the back-off/contention window parameter value. These linguistic rules shape the control knowledge (rule base) of the controller and describe how to best control the system, under differing operating conditions. Linguistic variables are used to describe the inputs and output of the fuzzy logic system, and retain linguistic values that describe specific characteristics of the variables. Linguistic values are generally descriptive terms such as "small", "large" and "very-large". Partitions are defined over the input and output space so as to establish the linguistic values characterizing the linguistic variables.

An offline intuitive tuning procedure is followed in order to obtain and set up the control rules and the values of the linguistic variables, following the successful, simple, and most commonly used procedure that was utilized in our earlier work [34]. The design objective is to maintain the controller as simple as possible to begin with, and only increase complexity, by adding more linguistic values, if required [34].

The basic idea following the rule base of the FLADS controller is two fold:

- Be aggressive when the traffic density of the channel is very high over the two successive sampling periods. In this case congestion starts to grow and fast relief is required.
- Be able to smoothly respond in the case of low density.

All other rules can represent intermediate conditions. Therefore, the constructed rule base provides the control mechanism with a highly dynamic accomplishment (see Table 2). Note that the actual number of rules implemented in the FLADS controller is reduced, since when the current channel traffic occupancy is "very-large", then the output control signal is always "very-large", irrespective of the status of the past channel traffic occupancy.

Table 2 FLADS linguistic rule base

CW(kT)	CTO(kT-T)					
		Z	VS	S	L	VL
CTO(kT)	*Z*	Z	Z	Z	VS	VS
	VS	VS	VS	VS	S	S
	S	S	S	S	S	S
	L	VL	L	L	L	L
	VL	VL	VL	VL	VL	VL

Table 2 notations for linguistic values: zero (Z), very-small (VS), small (S), large (L), very-large (VL)

5.3 Decision on the Membership Functions

The selection of the shapes of the membership functions of the linguistic inputs and output is open as many options are available (see, e.g. [20]). We have selected triangular shaped membership functions, in the FLADS control system, that offer computational simplicity. The selected membership functions representing the linguistic values for both the inputs and the output of the FLADS controller are shown in Fig. 2. Note that the same normalized membership functions are used for both of the controller's inputs. The same applies for the controller's output, where the dynamic value of the CW is obtained from the appropriate range [CW_{min}, CW_{max}] of each AC (as discussed in Sect. 2—Table 1). As seen in Fig. 2, we have used symmetric-and-equally spaced membership functions [34] that offer simplicity. The sum of the grade of membership of an input value, concerning the linguistic values of a specific input variable, is always one (see Eq. 2).

$$\sum_{k=1}^{m} \mu_k(x_i) \;=\; 1 \tag{2}$$

where $\mu_k(x_i)$ is the membership value of the input value x_i taken from the membership function of the linguistic value k, ($1 < k < m$, where m is the number of linguistic values of a linguistic variable), of the input variable of concern [20, 34].

This kind of design has as a result to have at most two membership functions overlapping, and as a consequence no more than four rules can be activated at a particular time. This offers computational simplicity on the implementation of the FLADS controller, which is a design objective.

The defuzzification method utilized by the FLADS controller is based on the well used "center of gravity" method [34, 35]; thus, the controller's output value is determined using Eq. (3):

$$CW_k = \frac{\int y\mu_C(y)dy}{\int \mu_C(y)dy} \tag{3}$$

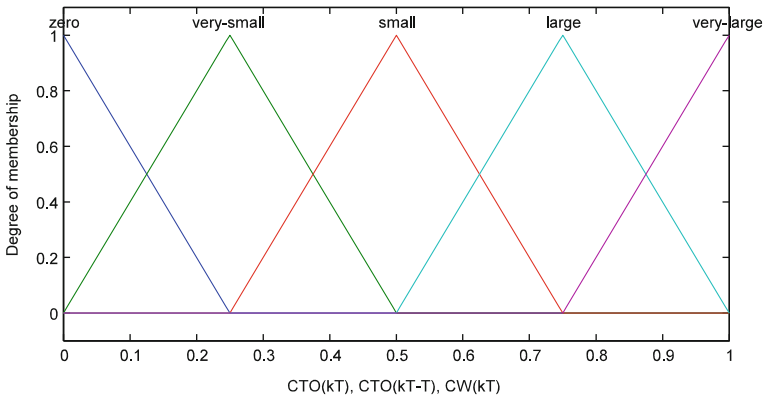


Fig. 2 Membership functions of the linguistic values representing the FLADS controller's input and output linguistic variables

where, $\mu_C(y) = \max(\mu_1(y), \mu_2(y), \ldots, \mu_N(y))$ is the membership degree of y in the aggregated fuzzy set C (which is found using the max-operation over all N implicated output fuzzy sets), and N is the number of linguistic rules. To further reduce computations, the output universe of discourse is discretized to m values, $Y = \{y_1, y_2, \ldots, y_m\}$, which gives the discrete fuzzy centroid (see Eq. 4). It is worth noting that the use of symmetric triangular membership functions makes the computation of the equation easy.

$$CW_k = \frac{\sum_{j=1}^{m} y_j \times \mu_C(y_j)}{\sum_{j=1}^{m} \mu_C(y_j)} \tag{4}$$

The resulted nonlinear control/decision surface of the FLADS controller can be seen in Fig. 3. This surface is formed based on the constructed rule base and the shapes of the linguistic values of both the inputs and output of the controller. The decision surface together with the rule base, as shown in Fig. 3 and Table 2, respectively, give the precise operation of the FLADS controller. The back-off/CW value is smoothly calculated, under the equilibrium state; that is, where the channel traffic occupancy is close to zero. On the other hand, the CW value increases harshly in the region beyond the equilibrium state, where congestion starts to grow and fast relief is required. Therefore, the decision support on when to access the communication broadcast channel, based on the dynamic CW value obtained, is expected to be more responsive than other conventional solutions (as for example, the 802.11p protocol), due to the human reasoning and the inbuilt nonlinearity.

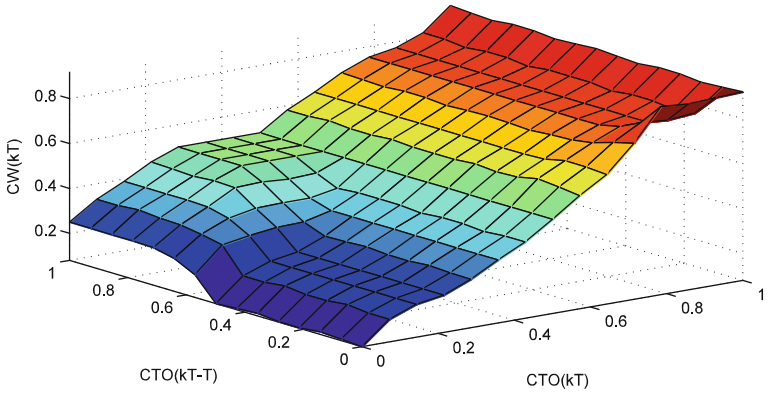


Fig. 3 Nonlinear control-decision surface of the fuzzy inference engine of FLADS controller

6 Performance Evaluation

An enriched simulative performance evaluation is used in order to exhibit the effectiveness and robustness of the fuzzy logic-based adaptive decision support scheme (FLADS), when applied in vehicular networks. A comparison is made with the IEEE 802.11p protocol. The performance is evaluated using the discrete event-based simulation environment OMNeT++ [36], which is a widely used network simulator, integrated by the Veins [37] open source framework for running vehicular network simulations.

The simulations study the behavior of the proposed controller and its counterpart when the broadcast communication channel is used in active road safety and control applications (including broadcast beaconing).

The common network parameters used in the simulations are:

- The channel data rate is set to 6 Mbps.
- Each autonomous vehicle broadcasts safety/control messages of 3200 bits size (as defined/used in [13]) every 5 ms. Thus, the generation rate is set to 200 Hz. Note that we have set different priorities, to create a mixture of differently prioritized types of traffic.
- The autonomous vehicles are randomly moving in a specified area, and they are all within the transmission range of each other.
- The sampling period is set to 20 ms (This is slightly bigger than the maximum back-off time that could be obtained, by having a contention window parameter of 1023 multiplied by the slot time of 13 μsec).
- The total simulation time is 15 s.

We investigate the performance of the tested schemes as the number of vehicles increases per simulation scenario, thus creating a very dense environment. This is done in order to monitor the behavior of the tested schemes when supporting the

Table 3 Network simulation parameters

Network parameter	Value
Channel data rate	6 Mbps
Message size	3200 bits
Generation rate/message interval	200 Hz/5 ms
Number of autonomous vehicles	4, 8, 16, 32, 64, 128
Vehicles movement	Yes—randomly
FLADS sampling period	20 ms
Total simulation time	15 s

channel access decision for vehicles in such scenarios. The number of vehicles is increased from 4 to 8, 16, 32, 64 and 128. A summary of the network simulation parameters used is shown in Table 3.

We have chosen the following important performance metrics for the evaluation of the tested schemes:

- Throughput—consequently the channel utilization
- Collisions
- Delay—the time needed to send a message from one vehicle to another
- Message reception rate—the number of messages received by a vehicle within a unit time period.

6.1 Throughput—Channel Utilization Study

As the number of vehicles increases (i.e. we have very high density) the proposed scheme outperforms the 802.11p in terms of throughput, and consequently the channel utilization (see Fig. 4). The fact that the FLADS controller provides better QoS than its counterpart is also evident in Fig. 5, where we show the channel utilization gain of the proposed scheme over the original 802.11p, as the number of vehicles increases. It is clearly shown that the FLADS controller can much better support the decision regarding the access of the broadcast communication channel as the vehicular environment gets dense. On the other hand, the 802.11p protocol behaves poorly in such conditions, and consequently, fails to support the strict requirements of the active road safety and control applications.

6.2 Collisions Study

The ability of a MAC protocol to control access to the communication channel, that is, to support the decision of the vehicles about when to access the channel, and consequently to minimize message collisions, is of prime importance in vehicular environments. This is especially true when active road safety and control

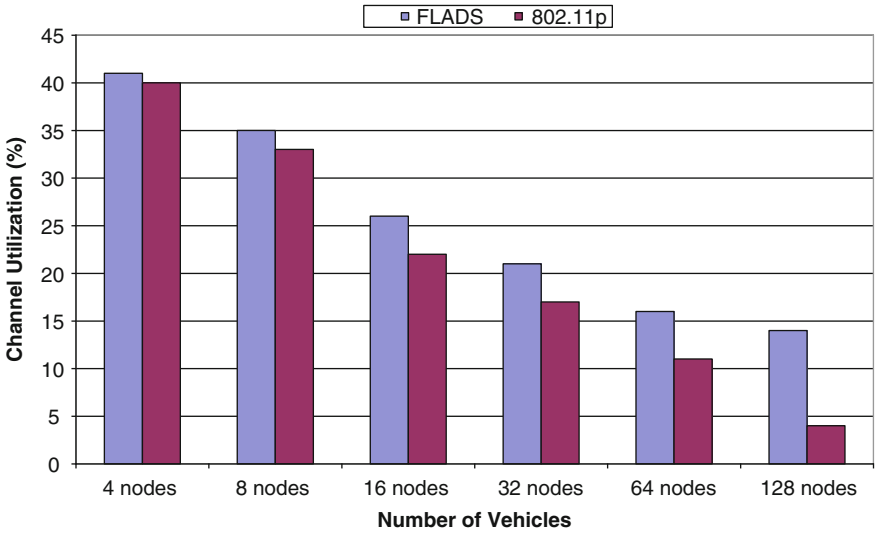


Fig. 4 Broadcast communication channel utilization

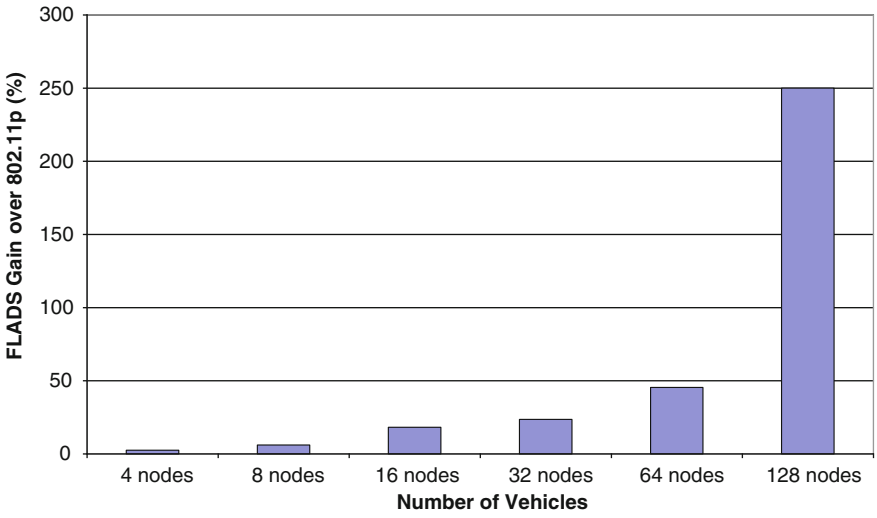


Fig. 5 Broadcast communication channel utilization: FLADS gain over 802.11p

messages are transmitted. We are interested in obtaining the Signal to Interference plus Noise Ratio (SINR), which is a good measurement of the quality of wireless connections, and hence can provide us with a collision model of the vehicular environment (that is obtained by the vehicular network simulator [37]). Taking this into consideration, we take the normalized average number of collisions in terms

of the collisions obtained when the FLADS controller is used, as the number of vehicles increases (see Fig. 6). It is clearly observed that the FLADS controller outperforms the original 802.11p protocol, as the conventional counterpart causes significant increase of the number of collisions as compared to the proposed controller. This can be further seen in Fig. 7, where the gain on reducing the collisions when the FLADS controller is used over the original 802.11p, is shown.

6.3 Delay Study

A significant performance metric obtained concerns the message delay; that is, the time needed to send a message from one vehicle to another. In [13], it is argued that the average delay is larger for higher CW values and increases with the number of vehicles. Also, based on the results in [13] the increase of the CW value did not yield the expected beneficial effects. However, as already discussed in Sect. 3, the work in [13] assumed that all vehicles in any scenario used the same CW settings. In our work, on the other hand, we offer the differentiation/prioritization of the various different applications that use the broadcast communication channel in terms of different minimum and maximum CW values (as defined in Table 1). Therefore, based on this fact, our network simulations can be seen as being closer to the actual vehicular environment. Due to the adaptive decision of the contention window value based on the dynamics of the environment, the FLADS controller manages to offer an average delay close to the one obtained by the conventional counterpart (see Fig. 8). Therefore, no significant overhead in terms of delay is added to the environment, due to the adaptiveness that is applied in the CW value as compared with the conventional scheme, where the CW value remains to its minimum value irrespective of the dynamics of the environment.

6.4 Message Reception Rate Study

The message reception rate is the number of messages received by a vehicle within a unit time period. This performance metric is very important as it denotes how many of the active road safety and control applications' messages are correctly received by the vehicles. Therefore, the higher the message reception rate, the greater the cooperative awareness among the vehicles. In Fig. 9, it is shown that the proposed controller outperforms its conventional counterpart in terms of average message reception rate, as the number of vehicles increases. With this performance metric, the superiority of the proposed controller to adequately support the decision on when to access the broadcast communication channel is evident. The FLADS controller manages to keep the average message reception rate at high levels even in the case of a dense environment, whereas its conventional counterpart fails to provide such a service to the applications.

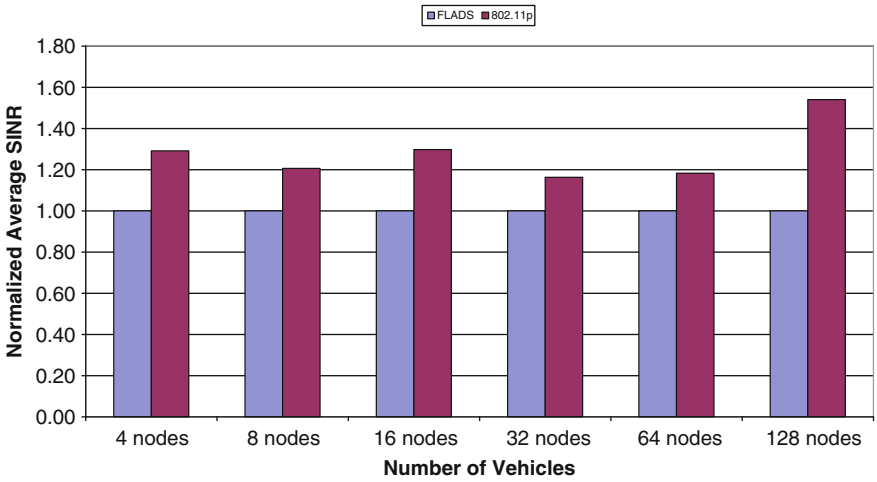


Fig. 6 Normalized average collisions

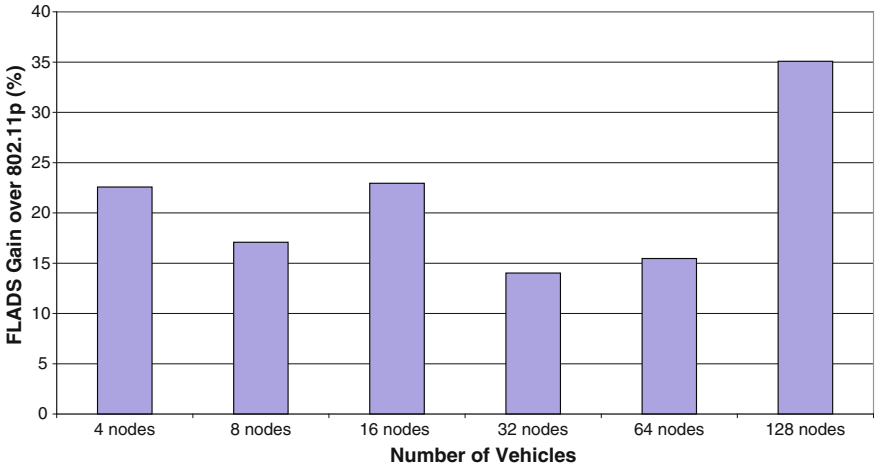


Fig. 7 Collisions gain of FLADS over 802.11p

6.5 Summary of Results

Overall, based on the obtained results, the proposed scheme outperforms the conventional one (i.e. the original 802.11p) taking into account all the measured performance metrics. Even though the drastic increase of the contention window

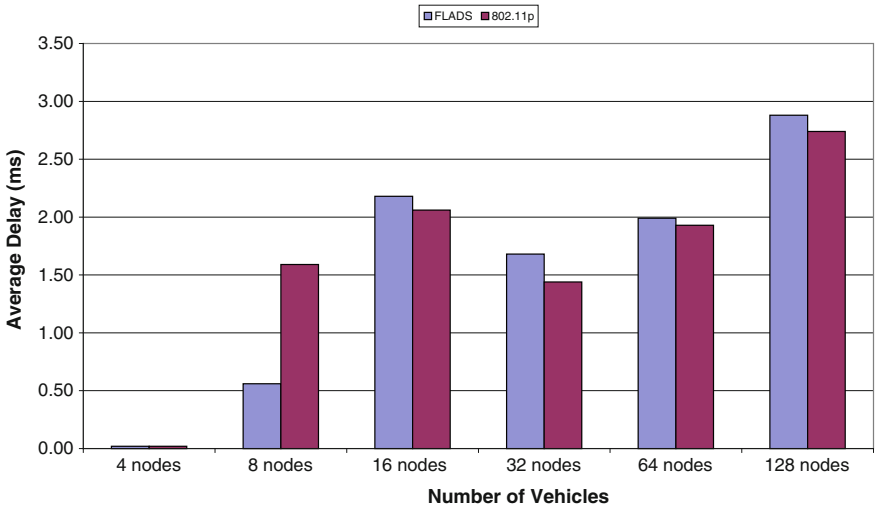


Fig. 8 Average delay

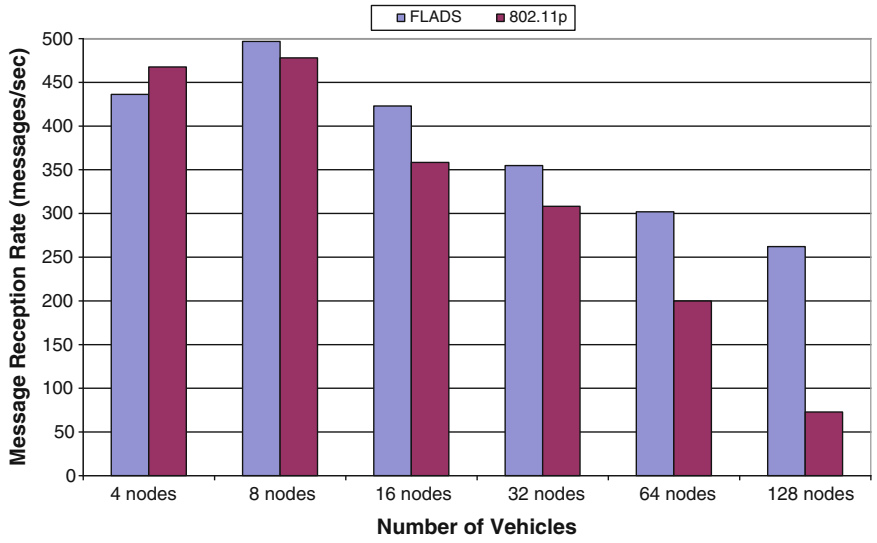


Fig. 9 Average message reception rate

can increase the delay, this is not the case in the proposed scheme. Due to the adaptive decision of the contention window size based on the dynamics of the environment, the FLADS controller manages to offer an average delay close to the

one obtained by the conventional counterpart. Further, as the number of vehicles increases, i.e. we have a very high density, the proposed scheme outperforms the 802.11p in terms of throughput, and consequently the channel utilization. As such, the proposed scheme outperforms the conventional counterpart in terms of messages reception rate. The same behavior is shown in terms of collisions.

From the results, it is clear that as the number of vehicles increases, there is a need to provide an intelligent decision support as to when the vehicles should try to access the broadcast communication channel. This is very crucial, especially for safety ITS applications, like vehicle collision and traffic jam avoidance.

7 Conclusions

It is our view that the use of a nonlinear control law for decision support in communication channel access can offer an efficient and effective control of the system. Furthermore, due to the high variability and dynamics in vehicular networks, a robust, intelligent controller is beneficial to keep the system in a controlled state, under differing conditions. Hence, given the need for such decision support methodology, we show the utility of fuzzy logic control to gather such objectives. Fuzzy logic control can be considered as suitable candidate for communication channel access decision support mechanism, since it is successfully utilized in numerous research fields to control nonlinear systems using linguistic information.

Specifically, an adaptive, intelligent controller, based on the traffic density, is presented to aid autonomous vehicles in deciding when to access the broadcast communication channel. This supports the cooperative awareness that is essential among autonomous vehicles, from which many applications can be depicted, under the active road safety and control applications, like reducing traffic jams and avoiding vehicle collisions. The results clearly show that the proposed fuzzy logic-based adaptive decision support solution outperforms the original 802.11p solution, in all the important performance metrics.

Of course many challenges about the communication channel access decision support remain open. For example, based on the dynamics of the vehicular environment, the transmission power level can be adjusted, as well as the message generation rate, so as to provide the necessary decision support. Thus, a more comprehensive study can be depicted to include not only the contention window adaptation, but the adaptation of other important network parameters (such as transmission power and message generation rate, as indicated above). With this approach, the influence of such parameters in the decision support in autonomous vehicular networks can be demonstrated.

References

1. G. Karagiannis, O. Altintas, E. Ekici, G. Heijenk, B. Jarupan, K. Lin, T. Weil, Vehicular networking: a survey and tutorial on requirements, architectures, challenges, standards and solutions. IEEE Commun. Surv. Tutorials **13**(4), 1–33 (2011)
2. IEEE Std P1609.1, Trial-use standard for wireless access in vehicular environments (WAVE)—resource manager, (2006) doi:10.1109/IEEESTD.2006.246485
3. IEEE Std P1609.2, Trial-use standard for wireless access in vehicular environments (WAVE)—security services for applications and management messages (2006), doi: 10.1109/IEEESTD.2006.243731
4. IEEE Std P1609.3, Trial-use standard for wireless access in vehicular environments (WAVE)—networking services (2007), doi: 10.1109/IEEESTD.2007.353212
5. IEEE Std P1609.4, Trial-use standard for wireless access in vehicular environments (WAVE)—multi-channel operation (2006), doi: 10.1109/IEEESTD.2006.254109
6. IEEE Std 802.11p, IEEE standard for information technology-telecommunications and information exchange between systems—local and metropolitan area networks—specific requirements Part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications—Amendment 6: Wireless access in vehicular environments 2010, Doi: 10.1109/IEEESTD.2010.5514475
7. D, Jiang, L. Delgrossi, IEEE 802.11p: towards an international standard for wireless access in vehicular environments. Paper presented at the 67th IEEE vehicular technology conference, VTC Spring 2008, Singapore, pp. 2036–2040, 11–14 May 2008
8. USA ITS Standards advisory, Dedicated Short Range Communications (DSRC). http://www.standards.its.dot.gov
9. D. Jiang, V. Taliwal, A. Meier, W. Holfelder, R. Herrtwich, Design of 5.9 GHz DSRC-based vehicular safety communication. IEEE Wireless Commun. **13**(5), 36–43 (2006)
10. R.A. Uzcategui, WAVE: a tutorial. IEEE Commun. Mag. **47**(5), 126–133 (2009)
11. A. Senart, M. Bouroche, V. Cahill, S. Weber, Vehicular networks and applications, in *Middleware for network eccentric and mobile applications*, ed. by B. Garbinato, H. Miranda, L. Rodriques (Springer, Heidelberg, 2009), pp. 369–381
12. E.M. van Eenennaam, K. Wolterink, G. Karagiannis, G.J. Heijenk, Exploring the solution space of beaconing in VANETs. Paper presented at the 1st IEEE vehicular networking conference (VNC2009), Tokyo, pp. 1–8, 28–30 Oct 2009
13. R. Reinders, M. van Eenennaam, G. Karagiannis, G. Heijenk, Contention window analysis for beaconing in VANETs. Paper presented at the 7th IEEE international wireless communications and mobile computing conference (IWCMC 2011), Istanbul, pp. 1481–1487, 4–8 July 2011
14. S. Eichler, Performance evaluation of the IEEE 802.11pWAVE communication standard. Paper presented at the 66th IEEE international vehicular technology conference (VTC-2007 Fall), Maryland, pp. 2199–2203, 30 Sept–3 Oct 2007
15. Y. Wang, A. Ahmed, B. Krishnamachari, K. Psounis, IEEE 802.11p performance evaluation and protocol enhancement. Paper presented at the IEEE international conference on vehicular electronics and safety, Columbus, Ohio, pp. 317–322, 22–24 Sept 2008
16. K. Bilstrub, E. Uhlemann, E.G. Strom, U. Bilstrup, On the ability of the 802.11p MAC method and STDMA to support real-time vehicle-to-vehicle communication. EURASIP J. Wireless Commun. Network. **2009**, 902414 (2009). doi:10.1155/2009/902414
17. C. Chrysostomou, C. Djouvas, L. Lambrinos, Applying adaptive, QoS-aware medium access control in priority-based vehicular ad hoc networks. Paper presented at the 16th IEEE symposium on computers and communications (IEEE ISCC 2011), Corfu, pp. 741–747, 28 June–1 July 2011

18. C. Chrysostomou, C. Djouvas, L. Lambrinos, Dynamically adjusting the min-max contention window for providing quality of service in vehicular networks. Paper presented at the 11th annual mediterranean ad hoc networking workshop (Med-Hoc-Net'12), Ayia Napa, pp. 16–23, 19–22 June 2012
19. L.A. Zadeh, Fuzzy Sets. Inf. control **8**, 338–353 (1965)
20. K. Passino, M. Yurkovich, *Fuzzy Control*, (Prentice Hall, USA, 1998) ISBN 0-201-18074-X
21. L.A. Zadeh, Outline of a new approach to the analysis of complex systems and decision processes. IEEE Trans. Syst. Man Cybernetics. **3**(1), 28–44 (1973)
22. E.H. Mamdani, Applications of fuzzy algorithms for control of simple dynamic plant. Proc. IEE **121**(12), 1585–1588 (1974)
23. A. Pitsillides, A. Sekercioglu, Congestion control, in *Computational intelligence in telecommunications networks*, ed. by W. Pedrycz, A.V. Vasilakos (CRC Press, Boca Raton, 2000), pp. 109–158. ISBN 0-8493-1075-X
24. A. Sekercioglu, A. Pitsillides, A. Vasilakos, Computational intelligence in management of ATM networks. J.Soft Comput. **5**(4), 257–263 (2001)
25. B. Azvine, A. Vasilakos, Application of soft computing techniques to the telecommunication domain, ed. by G. Tselentis (ERUDIT Roadmap, 2000), pp. 89–110
26. E. Morales, M. Polycarpou, N. Hemasilpin, J. Bissler, Hierarchical adaptive and supervisory control of continuous venovenous hemofiltration. IEEE Trans. Control Syst. Technol. **9**(3), 445–457 (2001)
27. A. Sekercioglou, A. Pitsillides, G.K. Egan, Study of an adaptive fuzzy controller based on the adaptation of relative rule weights. Paper presented at the 2nd ANZIIS, Brisbane, pp. 204–208, 29 Nov–2 Dec 1994
28. A. Pitsillides, A. Sekercioglou, G. Ramamurthy, Effective control of traffic flow in ATM networks using fuzzy explicit rate marking (FERM). IEEE J. Sel. Areas Commun. **15**(2), 209–225 (1997)
29. C. Douligeris, G. Develekos, A fuzzy logic approach to congestion control in ATM networks. Paper presented at the of IEEE ICC'95, Seattle, pp. 1969–1973, 18–22 June 1995
30. L. Rossides, A. Sekercioglu, S. Kohler, A. Pitsillides, T.G. Phuoc, A. Vassilakos, Fuzzy logic controlled RED: congestion control for TCP/IP diff-serv architecture. Paper presented at the 8th European congress on intelligent techniques and soft computing (ESIT2000), Aachen, pp. 263–269, 14–15 Sept 2000
31. L. Rossides, C. Chrysostomou, A. Pitsillides, A. Sekercioglu, Overview of Fuzzy-RED in Diff-Serv Networks, in *Proceedings of Soft-Ware 2002*, vol. 2311, Lecture notes in computer science, ed. by D. Bustard, W. Liu, R. Sterritt (Springer, Heidelberg, 2002), pp. 1–13
32. C. Chrysostomou, A. Pitsillides, L. Rossides, M. Polycarpou, A. Sekercioglu, Congestion control in differentiated services networks using Fuzzy-RED. Special issue on control methods for telecommunication networks, IFAC Control Eng. Pract. J. **11**(10), 1153–1170 (2003)
33. C. Chrysostomou, A. Pitsillides, A. Sekercioglu, Fuzzy explicit marking: a unified congestion controller for best-effort and diff-serv networks. Elsevier Comput. Netw. J. **53**(5), 650–667 (2009)
34. C. Chrysostomou, Fuzzy logic based AQM congestion control in TCP/IP networks. PhD thesis, University of Cyprus (2006), http://www.netrl.cs.ucy.ac.cy/images/thesis/chrysostomou-phd-thesis-sep06.pdf
35. E.H. Mamdani, S. Assilian, An experiment in linguistic synthesis with a fuzzy logic controller. Int. J. Man Mach. Stud. **7**(1), 1–13 (1975)
36. OMNeT++ discrete event-based simulation framework. http://www.omnetpp.org
37. C. Sommer, R. German, F. Dressler, Bidirectionally coupled network and road traffic simulation for improved IVC analysis. IEEE Trans. Mob. Comput. **10**(1), 3–15 (2011)

Adaptive Trajectory Tracking of Wheeled Mobile Robot with Uncertain Parameters

Kanwal Naveed, Zeashan H. Khan and Aamir Hussain

Abstract A wheeled mobile robot (WMR) belongs to the class of non-holonomic systems with highly nonlinear dynamics. Because of their fast maneuvering and energy saving characteristics, these robots are especially popular in following or tracking a pre-defined trajectory. The trajectory of a WMR is controlled with the help of two very different control schemes namely model dependent approach and model free approach. While the model dependent approach relies on a particular model for the controller design, the model free method controls the trajectory with the help of learning methods. A Direct Model Reference Adaptive Controller (D-MRAC) is described for the model based technique, while an Adaptive Neuro-Fuzzy Inference System (ANFIS) is used for the model-free adaptive control design. With the help of simulations, it is shown that data driven intelligent approach is comparable to model dependent approach in terms of tracking performance and therefore can be preferred over complex model dependent adaptive algorithms.

Keywords Mobile robots · Adaptive control · Artificial intelligence · Trajectory tracking

1 Introduction

For many industrial applications, finding a model from physical laws that is both simple and reliable for control design is a tough undertaking. When a set of measurements is available; the control law can be computed from data without relying on knowledge of the underlying physics. Specifically, in "model-based" approaches, a model of the system is first derived from data and then a controller is computed based on such a model. In "data driven" approaches, the controller is

K. Naveed (✉) · Z. H. Khan · A. Hussain
National University of Science and Technology (NUST), Islamabad, Pakistan
e-mail: kanwal_naveed41@yahoo.com

Z. H. Khan et al. (eds.), *Computational Intelligence for Decision Support in Cyber-Physical Systems*, Studies in Computational Intelligence 540,
DOI: 10.1007/978-981-4585-36-1_8,

directly derived from experimental data. It is common belief that finding a good model of the plant is always the best way towards controller design; therefore model-based approaches are always preferred. This chapter presents some interesting results to show that the data-driven solutions might sometimes be preferable and perform comparable performance as compared to model based approaches, thus motivating research in these design methods, which are less time-consuming and insensitive to modeling errors. Some recent improvements of the weak points of such techniques are also discussed to lay the foundations for future research.

In this chapter, the problem of trajectory tracking of wheeled mobile robots is considered. These systems are classified as unmanned ground vehicles with the ability of moving around any area for the purpose of surveillance and sensing. They are also capable of carrying out tasks which are either too dangerous or tiresome for human beings among which the most popular are industrial and search/petrol applications. They can be grouped in a communication environment where these autonomous systems can share data between themselves for cooperative decision to form a robotic cyber physical system. These types of applications require WMR to follow some kind of pre-defined trajectory. In general, robots are not capable of taking decisions on their own because they have limited computational power and intelligence embedded onboard. In order for them to carry out different tasks, these mobile platforms must have at least four parts: Sensors, effectors, actuators and controllers. In case of WMR, the role of effectors is played by the wheels. The most important sub-system of these is the controller, which, with the help of sensors enables the robot to achieve its goals. In case of trajectory tracking, the controller is designed in such a manner which empowers WMR to reach any predefined track.

Many radical methods projected for trajectory tracking control uses the dynamic model of the WMR. In case of linear control design, a correct dynamic model and precise parameters of the system is required. However, in reality, the dynamic model of WMR may face some amount of inaccuracy in parameter identification or there may be environmental variations in dynamic model parameters. The inaccuracy and variations result in trajectory tracking errors, and may render the system unstable in some cases. Overcoming these errors and uncertainties in the dynamic model of WMR has been the subject of interest to researchers over a past few years. Many efficient control schemes have been introduced for this purpose which includes robust control, H-infinity control and adaptive control (Fig. 2).

However most of these techniques offer complicated control system design based on nonlinear model of WMR which may lead towards implementation difficulties. This fact highlights the need for a system that will not only overcome all these problems but is also easier to implement. This task can be achieved through a “model-free system” which will add some intelligence to the WMR thereby enabling it to overcome these uncertainties and disturbances. The supplement of natural intelligence in WMR is accomplished with the help of Artificial Intelligence [1]. Over a past few years, a lot many variations and advancements have been introduced in the field of Artificial intelligence. A very popular among these is Adaptive Neuro-Fuzzy Inference System (ANFIS).

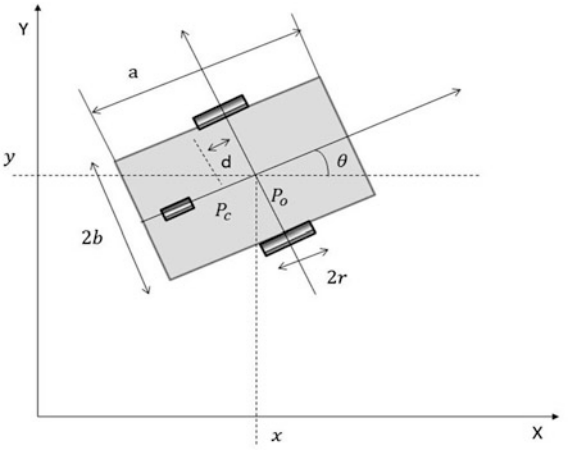


Fig. 1 Configuration of WMR

It is usual to think that if something is "*intelligent*" then it should be able to "*adapt*" to its environment and thus one may use adaptive controller. However, Artificial intelligence and Adaptive control are two different techniques. An adaptive controller has some built in algorithmic technique for automatic adjustment of controllers in real time, in order to achieve/maintain a desired level of control system performance even when the parameters of the plant dynamics are unknown and/or change in time [2]. The main purpose of this chapter is to discuss and compare trajectory control performance by using model-based as well as model-free adaptive control algorithms.

2 Model of Wheeled Mobile Robot

Wheeled mobile robot belongs to the class of movable autonomous platforms. Among many types of WMR differential drive robot is selected for this chapter because of its simple structure, light weight and ability to spin at a place. The complete model of WMR comprises of the kinematic and dynamic model as follows.

Kinematic model: It is based on the assumption that the robot moves on a flat surface with no side slip [3]. Linear velocity v_b and angular velocity w_b of WMR is responsible for controlling its motion. Using these velocities the kinematics of WMR can be defined by following relation:

$$\dot{q} = Gv(t) = \begin{bmatrix} \cos\theta & 0 \\ \sin\theta & 0 \\ 0 & 1 \end{bmatrix} \tag{1}$$

where $v(t) = [v_b w_b]^T$

Table 1 Parameters of WMR

Parameters	Description
r	Left and right wheel radius
$2b$	Distance between left and right wheels
d	Distance between P_c and P_o
m_c	Platform mass of WMR without wheels
m_w	Each driving wheel plus motor rotor mass
I_c	Platform moment of inertia without wheels
I_w	Each wheel and the motor rotor moment of inertia about the wheel axis
P_o	The center of mass of WMR

Dynamic model: It is derived using the Lagrange equations instead of Newton mechanism order to avoid complexity. After implementing the Lagrange equations and incorporating the motor actuator dynamics [4] the complete dynamic model of the WMR can be written as:

$$\bar{M}\dot{v}(t) + \bar{F}(\dot{q}).v(t) = K_1\bar{B}.u_c \tag{2}$$

where, $\bar{M} = G^T MG$, $\bar{F} = G^T M\dot{G} + G^T FG + K_2 BW$, $\bar{B} = G^T B$

$$M(q) = \begin{bmatrix} m & 0 & m_c d\sin\theta \\ 0 & m & m_c d\cos\theta \\ m_c d\sin\theta & -m_c d\cos\theta & I \end{bmatrix}$$

$$F(q,\dot{q}) = \begin{bmatrix} 0 & 0 & m_c d\dot{\theta}\cos\theta \\ 0 & 0 & m_c d\dot{\theta}\sin\theta \\ 0 & 0 & 0 \end{bmatrix}$$

$$B(q) = \frac{1}{r}\begin{bmatrix} \cos\theta & \cos\theta \\ \sin\theta & \sin\theta \\ b & -b \end{bmatrix} \tag{3}$$

where $m = m_c + 2m_w$ and $I = I_c + 2I_m + m_c d^2 + 2m_w b^2$ are defined in Table 1, with reference to Fig. 1, and where M(q) is an n × n inertia matrix, F(q) is a n × n matrix which represents the centripetal and coriolis forces, B(q) is an n × (n − m) input matrix.

3 Trajectory Tracking of Wheeled Mobile Robot

The incorporation of wheels in a robot makes it much more reliable for applications which involve Trajectory tracking. Whether a robot has to work in an industry or carry out the surveillance tasks it requires the knowledge of some predefined path or track to be followed. The predefined trajectory helps the robot in following the correct direction in a right manner. While talking about trajectory

tracking it should not be forgotten that in the task of tracking a trajectory an extra time constraint is also involved. In standard terminology, the trajectory tracking problem can be stated as a time parameterized control problem, in which the geometric path has an associated timing specification. Thus, the control objective of trajectory tracking is considered to deal with a time $\times$ space task, which is the intersection set of spatial task and temporal task [5].

With the above problem stated, it is now clear that controlling the trajectory of a robot is very important, not only to keep the robot on a defined track but to reach the target(s) within the specified time limit. A Wheeled Mobile Robot delivering certain items in an industry requires a track which it can follow in order to drop them at the specified target within a precise time. A little variation in the track/timing of robot can create enormous problems in the item deliveries. Moreover, same problems will occur if the Wheeled Mobile Robot has to perform the surveillance job. If it fails to follow the defined track then this blunder can bring many security hazards. Therefore, the importance of controlling the trajectory of the robot has attracted the attention of researchers for a past few years. Many different types of control schemes are available for the purpose of trajectory tracking of WMR, which can be classified in two distinct types:

Model based control methods: These methods involve complete knowledge of system/plant kinematics and dynamics before the controller can be designed. The procedure of controller design depends greatly upon the plant parameters and requirements.

Model-free control methods: These methods are called model free because the plant can be controlled without making any assumption about the structure or nature of plant (Wheeled Mobile Robot) being controlled or its environment.

The former is much older as compared to the later type. A great number of linear controllers have been developed for the tracking control of Wheeled Mobile Robot. However, advancements in the model free approach are also in progress as its simplicity and ease of usage is attracting the researchers all over the world. Later sections discuss the trajectory tracking control of Wheeled Mobile Robot using these two methods in detail.

4 Model Based Control Methods

Design and development of controllers via mathematical model of a system/plant is a very frequent approach. Model based design is a mathematical method of handling problems which are related to complex control design. Controller which is designed in the light of plant model is usually obtained after going through three distinct phases:

Plant model: In model based approaches model of the plant under operation is necessary to obtain. Without the knowledge of plant parameter particulars it is impossible to design a controller. The model of the plant can be obtained through data-driven approaches or by using Newton or Lagrange rules. Both the

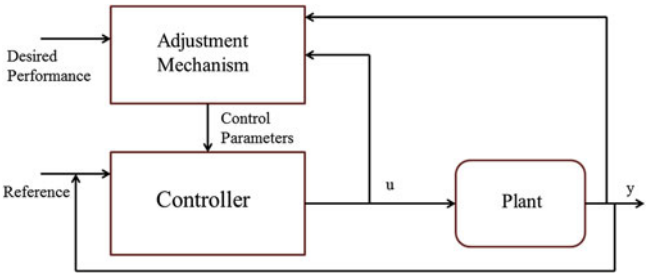


Fig. 2 Structure of adaptive controller

approaches are equally useful and popular. The utility of any approach depends upon the feasibility of conditions.

Controller design: After the plant model is acquired and the user has complete knowledge of plant behavior and its parametric characteristics, the controller can be designed using this plant model. Design of a controller depends upon the type of application and the precision and accuracy it requires. More the need of accuracy more complex is the design of the controller. A number of linear and nonlinear techniques are available now-a-days which help in the design of the required controller.

Simulation: After the controller is successfully designed, the next step is to test the efficiency of the controller. This can be achieved via any useful simulation tool. Any type of modeling or design errors can simply be identified through the simulations.

Although, the model based control approach requires several steps in order to design a controller, which can suffice the needs of the user, it is very popular among researchers. This is because this approach has been practiced over many years and a number of controller designs are available for the users, which can easily meet their conditions. Such types of controllers include conventional linear controllers for example Proportional Integral Derivative (PID) controllers. However, these controllers are rendered useless under uncertainties and disturbances. Under such circumstances, adaptive controllers are especially helpful with their ability to update controller gains. However, in the presence of stochastic disturbances and variations in plant parameters these controllers fail to provide an optimum performance. The main reason behind this failure is that these controllers require priori information about the plant parameters in order to design their gains. Even a little variation in plant (robot) parameters can lead to degraded performance of these controllers.

The solution to this problem has been answered by adaptive controllers which include an adjustment mechanism to update the parameters of either the plant or the controller (Fig. 2). The adjustment mechanism derives an update law for the parameters and in this way the control (or plant) parameters are being updated. This update quality enables the adaptive controllers to overcome any type of parametric uncertainty. But this adjustment mechanism is also responsible for

making these controllers non-linear thereby making it difficult to implement. However, due to the advantages provided by these controllers many researchers are being attracted towards them. A number of studies have been performed to examine adaptive control for robotics [3, 4, 6, 7].

Adaptive controllers come in a variety of types among which the most popular is the Model Reference Adaptive Control (MRAC). This is because of its simple nature of directly taking error (between desired and actual plant outputs) into the account in order to achieve the control task. This chapter considers MRAC as a Model Dependent method in order enable WMR follow the desired trajectory in presence of parametric uncertainty.

4.1 MRAC Based Trajectory Tracking

Model Reference Adaptive Control (MRAC) is a closed loop control system which takes in account a reference model in order to yield the desired output from plant. This reference model when given the reference input provides us with the desired output. The reference model can either be linear or nonlinear and is designed according to user specifications e.g. transient response time. This type of adaptive control generates the update law for the parameters of controller based upon the error between the desired output and the actual plant output.

MRAC uses two different types of methods to update the control parameters:

- MIT rule
- Lyaponuv Stability approach.

The most popular among the two is the latter. Model reference adaptive control comes in two different types:

1. Direct model reference adaptive control: Directly takes in account the error between the desired output and the actual plant output to update the control parameters. It does not include any parameter estimation criterion in case of plant parameter perturbation. This chapter will focus on direct MRAC.
2. Indirect model reference adaptive control: Plant parameters are first estimated using an identification model and then the control update mechanism is applied (Fig. 3).

4.2 Controller Design

For the trajectory tracking task of wheeled mobile robot (WMR) a direct model reference adaptive control scheme is devised which will overcome the plant parametric uncertainty and disturbances. The complete model of WMR has already been described in section. It is being assumed that the kinematic model of

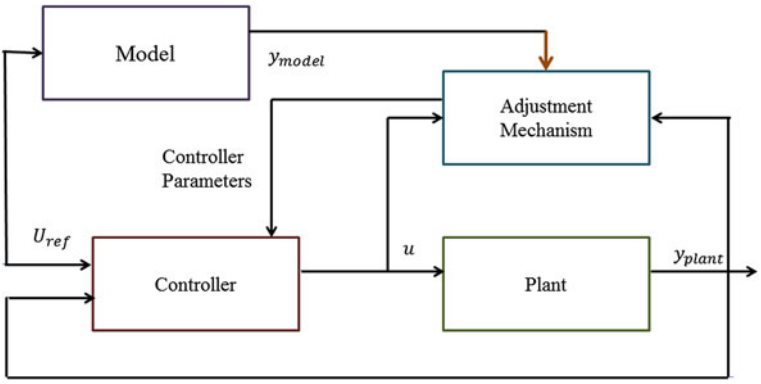


Fig. 3 Model reference adaptive control block diagram

the WMR is well established and only the dynamic model of WMR contains uncertainties therefore the MRAC will be incorporated in dynamics of WMR. The controller design for trajectory tracking of WMR incorporates two stages (Fig. 4):

- Kinematic controller
- Dynamic controller.

4.3 Dynamic Controller

As it can be observed that dynamic model is nonlinear in nature and includes time varying and uncertain terms:

$$\bar{M}(q)\dot{v}(t) + \bar{F}(\dot{q})v(t) = u_b \tag{4}$$

where $u_b = K_1\bar{B}u_c$, therefore a direct model reference control scheme is developed in order to achieve trajectory tracking of WMR. The convergence of this arrangement necessitates that the WMR model is slowly time varying and close to the desired model. However, the second condition can be relaxed by taking benefit of any known dynamics [3]. The dynamics of WMR can be separated in terms of known and unknown dynamics as:

$$\bar{M} = \bar{M}_k + \bar{M}^* = (\bar{M}_k - I) + \left(I + \bar{M}_u^*\right) = (\bar{M}_k - I) + \bar{M}_u$$

$$\bar{F} = \bar{F}_k + \bar{F}_u \tag{5}$$

The subscript ‘k’ and ‘u’ denotes known and unknown parameters. The known dynamics are forwarded to feedback linearization to calculate the linear constant

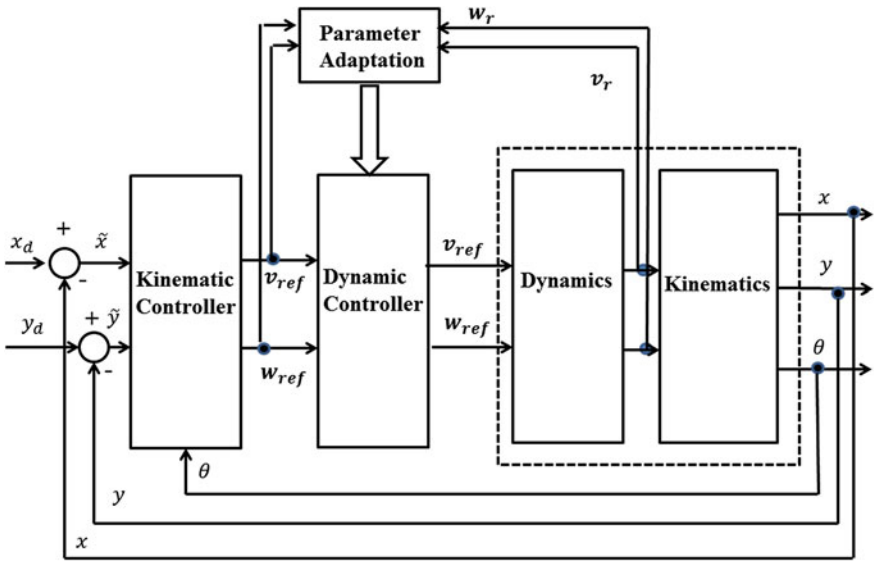


Fig. 4 Controller architecture for direct MRAC

controller gain. Note that if there is no parametric uncertainty then $\bar{M}_u = I(\bar{M}_u^* = 0)$ and $\bar{F}_u = 0$. Substituting (5) in (4) we get:

$$\bar{M}(q)\dot{v}(t) + \bar{F}(\dot{q})v(t) = u_b' \tag{6}$$

$$u_b' = u_b - (\bar{M}_k - I)\dot{v}(t) - \bar{F}_k(\dot{q})v(t) \tag{7}$$

If u_b' is known then from Eq. (8) u_b can be recovered therefore:

$$u_b = u_b' + (\bar{M}_k - I)\dot{v}(t) + \bar{F}_k(\dot{q})v(t) \tag{8}$$

After the removal of known dynamics by transformation of (7) the control problem for (4) can simply be stated as that of designing u_b' to obtain u_b such that the actual velocities of WMR follow the desired velocities in order to achieve the task of trajectory tracking. Defining $x = v$ we get:

$$\dot{x}(t) = Ax(t) + Bu_b' \tag{9}$$

where, $A = -\bar{M}_u^{-1}\bar{F}_u$ and $\bar{B} = \bar{M}^{-1}u$. The reference model with the desired stated and the reference input can be written as:

$$\dot{x}_d = A_r x_d + B_r u_{ref} \tag{10}$$

Here $A_r = -K$ and $B_r = I$ with K being the diagonal matrix. K_{11}, K_{22} Are constant terms equal to α_i which is inversely proportional to the time constant of desired transient behavior of the system. The direct model reference adaptive control strategy will make an effort to make the error between the reference and

actual states asymptotically zero. In this way, the adaptive scheme will drive the WMR to follow desired trajectory. Defining $x_d = v_d$ the error between the reference and actual states takes the following form:

$$e = x_d - x \tag{11}$$

Dynamics of error are being obtained by taking derivative of above equation and making use of Eqs. (9) and (10):

$$\dot{e} = A_r e + (A_r - A)x - \bar{B}u_b^{'} + B_r u_{ref} \tag{12}$$

The reference input is given as:

$$u_{ref} = \dot{v}_d + Kv_d \tag{13}$$

The control law can be written in the following form:

$$u_b' = u_l + u_a \tag{14}$$

This controller has two parts. The linear feedback controller derived from the known dynamics is given as follows:

$$u_l = -Kx + u_{ref} \tag{15}$$

And the adaptive controller devised for the unknown part is given as:

$$u_a = -\gamma x + \gamma_r u_{ref} \tag{16}$$

Equation (14) includes two gains which play important role in the design of adaptive controller. These gains are not constant but are incessantly updated such that the error approaches zero asymptotically. The update law is being provided by the Lyapunov stability criterion. Replacing (14) in (12) the error dynamics becomes:

$$\dot{e} = A_r e + C_e + D_e u_{ref} \tag{17}$$

$$C_e = \bar{M}_u^{-1}(\bar{F}_u + K + \gamma) - K \tag{18}$$

$$D_e = I - \bar{M}_u^{-1}(I + \gamma_r) \tag{19}$$

Theorem 1. Using P as a positive definite result of Lyapunov equation with $Q > 0$:

$$A_r^T P + PA_r = -Q \tag{20}$$

If the filtered error is defined as:

$$we = Pe \tag{21}$$

And the adaptive gains in (16) are updated such as:

$$\dot{\gamma} = -\bar{a}wev^T,\ \dot{\gamma}_r = \bar{b}weu_{ref} \tag{22}$$

where $\bar{a}$ and $\bar{b}$ are positive scalars. Thus by utilizing (22) in the control law demonstrated from (14) to (16) asymptotic stability of closed-loop error in (11) is guaranteed.

Proof: Choose the following Lyapunov candidate function using $P > 0$:

$$L = \frac{1}{2}\left[e^T Pe + \frac{1}{\bar{a}} tr\left(C_e^T Y C_e + D_e^T D_e\right) + \frac{1}{\bar{b}} tr(D_e^T Y D_e - D_e^T D_e)\right] \tag{23}$$

After differentiating L and using some common mathematics properties we obtain:

$$\dot{L} = \frac{1}{2}\left(e^T\left(A_r^T P + PA_r\right)e\right) + tr\left(C_e^T\left(Pex^T + \frac{Y\dot{C}_e}{\bar{a}}\right)\right) + tr\left(D_e^T\left(Peu_{ref}^T + \frac{Y\dot{D}_e}{\bar{b}}\right)\right) \tag{24}$$

In order to make $\dot{L}$ negative definite, we can make first term negative and other two terms identically equal to zero. This is possible if P is chosen from (20) and by choosing adaptation laws as in (22). Taking derivative of (18) and (19) assuming that plant parameters are changing slowly we obtain:

$$\dot{C}_e = \bar{M}_u^{-1}\dot{\gamma} \tag{25}$$

$$\dot{D}_e = -\bar{M}_u^{-1}\dot{\gamma}_r \tag{26}$$

where

$$\dot{\gamma} = -\bar{a}wev^T \tag{27}$$

$$\dot{\gamma}_r = \bar{b}weu_{ref} \tag{28}$$

Therefore, the Eq. (24) turns into:

$$\dot{L} = -\frac{1}{2}e^T Qe \tag{29}$$

From (29), it is easily deduced that 'L' is decreasing constantly. This proves that 'e' approaches to zero asymptotically. Equations (27) and (28) provides us with the update rule for the adaptive part of the controller. However if the system has no uncertainty ($\bar{M}_u = I$, $\bar{F}_u = 0$, $\gamma = 0$, $\gamma_r = 0$) then (12) is reduced to the following form:

$$\dot{e} = A_r e \tag{30}$$

Consequently, the only task is to choose a positive diagonal matrix A_r so that the asymptotic stability of error is guaranteed.

4.4 Kinematic Controller

The basic kinematic model of WMR is exploited to develop the kinematic controller. The sole purpose of this controller is to provide desired velocity profiles to the dynamic controller. The controller is established using error between the desired and the actual trajectory and is given by [8].

$$\begin{bmatrix} v_d \\ w_d \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \dot{x}_d + n_x \tanh \dfrac{m_x}{n_x}\tilde{x} \\ \dot{y}_d + n_y \tanh \dfrac{m_y}{n_y}\tilde{y} \end{bmatrix} \tag{31}$$

Here $\tilde{x} = x_d - x$, $\tilde{y} = y_d - y$ are the errors between the actual and the desired position. Moreover m_x and $m_y > 0$ are the controller gains and n_x, $n_y > 0$ are saturation constants.

4.5 MRAC Performance Results

In order to evaluate the performance of Model reference adaptive controller for nominal and parametric uncertainty cases for trajectory tracking of Wheeled Mobile Robot two different trajectories are considered. The performance is evaluated by overlapping the results of nominal and uncertainty case. The physical and design for MRAC parameters for computer simulations are obtained from [3] (Figs. 5, 6, 7, 8):

$$\begin{aligned} r = 0.15\,\text{m}, b = 0.75\,\text{m}, d = 0.3\,\text{m}, m_c = 36\,\text{kg}, m_w = 1\,\text{kg}, I_m = 0.0025\,\text{kg.m}^2, \\ I_c = 15,625\,\text{kg.m}^2, K_1 = 7.2, K_2 = 2.59, m_x = m_y = 10, \\ n_x = n_y = 4, \bar{a} = \bar{b} = 100, K = diag(1000, 1000), K_x = 9, K_r = 40. \end{aligned}$$

5 Model Free Control Approach

As opposed to the model dependent approach, this method is independent of the physical model of the plant. These methods incorporate learning approaches and/or logic statements in order to carry out control operation. Being able to operate without the plant model knowledge makes this approach independent of the model uncertainties. Such techniques are most widely used technique for decision support in robotics. Model free approach usually incorporates Artificial Intelligence in order to carry out the control operation. AI enables the robots to learn just about anything, permits them to reason and formulate original ideas. However, this level

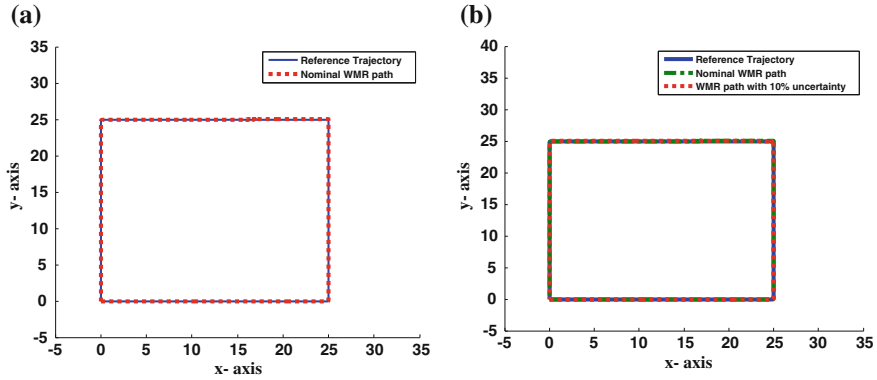


Fig. 5 Square trajectory tracking using MRAC. **a** Nominal plant model, **b** 10 % uncertainty

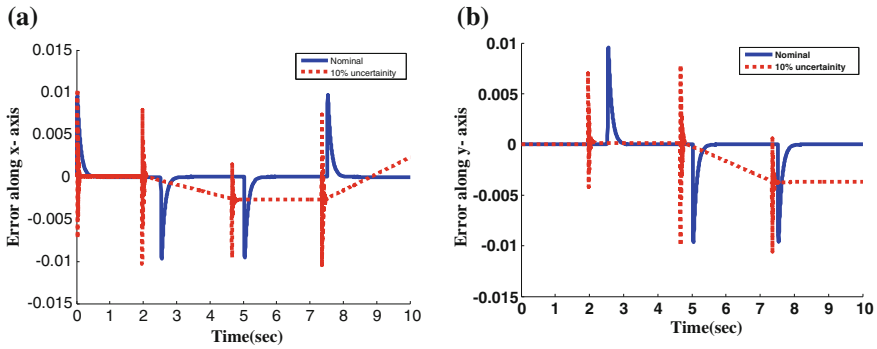


Fig. 6 Trajectory tracking error using MRAC. **a** x-axis, **b** y-axis

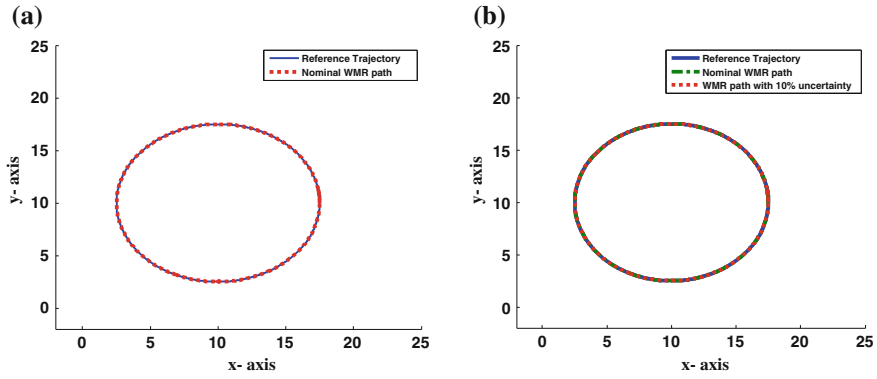


Fig. 7 Circular trajectory tracking using MRAC. **a** Nominal plant model, **b** 10 % uncertainty

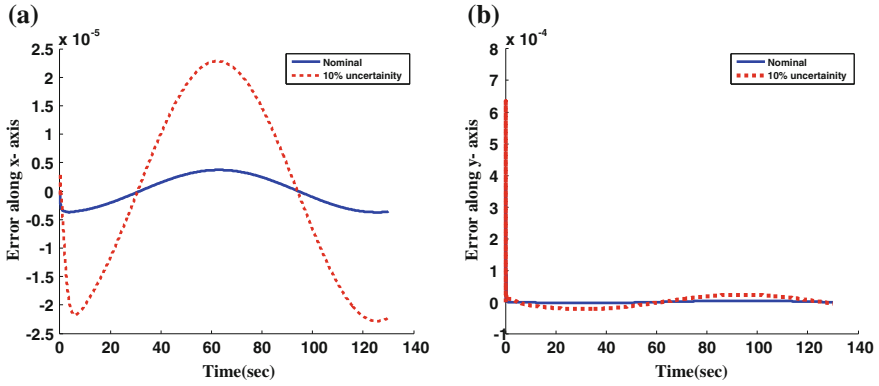


Fig. 8 Trajectory tracking error using MRAC. **a** x-axis, **b** y-axis

of AI is not yet attained but a lot more improvement has been made with more restricted AI. Currently, AI robots can imitate certain precise essentials of intellectual aptitude. Artificial intelligence works with the help of Artificial Neurons (Artificial Neural Network) and scientific theorems (If–Then Statements, Logics). AI is in reality "model free" rather it is data driven. It does not require any model to come up with the output. This quality of AI is especially helpful in case of highly nonlinear dynamical nature of WMR. The most popular among these is Adaptive Neuro Fuzzy Inference System (ANFIS).

5.1 Adaptive Neuro Fuzzy Learning Based Trajectory Tracking

Adaptive Neuro-fuzzy inference system has the advantage of neural networks learning which a model free approach for trajectory controller design. It incorporates the best features of both the above described techniques. The Fuzzy inference system (FIS) it represents the a priori knowledge in the form of fuzzy sets which reduces the optimization search space. The Neural Network part accepts the back propagation algorithm to structure network by learning from data.

The WMR model described above, is only required for the Model Reference Adaptive Control. This section requires only training on datasets without any model of WMR. This is because ANFIS has the ability to approximate the function by using the neural networks. All an ANFIS require is an appropriate training set in order to achieve its target. However, while performing the training over or under training must be avoided. The best results can only be obtained if an optimum training set is chosen.

5.2 ANFIS Network

Adaptive Neuro-fuzzy system utilizes a hybrid-learning rule to optimize the fuzzy system parameters of a first order Sugeno system and provides us with only a single output. The basic ANFIS network for trajectory tracking of WMR consists of following structure (Fig. 9):

The above ANFIS consists of two inputs and a single output. The output represents the trajectory of WMR and the inputs represent the Reference trajectory and the parameter uncertainty. In case if all the parameters are known then ANFIS can work perfectly with just the single input this will be the reference trajectory in that case. Following is the layer-by-layer explanation of ANFIS structure [9]:

Layer 1: This layer is actually used to generate the membership functions for the given inputs. In our case the chosen a single membership function is being chosen which, is a bell shaped gauss function. The output of layer 1 can be written as:

$$O_1 = h(x) \tag{32}$$

where 'h' represents the membership function.

Layer 2: This layer is used to produce the firing strengths for rules which are either in form of products or T-norms:

$$O_{2,i} = w_i \prod_{k=1}^{m} h(x),\ i = 1 \text{ to } 5 \tag{33}$$

Layer 3: The task of normalizing the strengths or weights is carried out in this layer:

$$O_{3,i} = \bar{w}_i = \frac{w_i}{w_1 + w_2 + w_3 + w_4 + w_5} \tag{34}$$

Layer 4: This layer produces the single output of ANFIS network:

$$O_4 = y = \sum_i \bar{w}_i f_i = \sum_i \bar{w}_i (p_i x + q_i x + r_i) \tag{35}$$

5.3 SDRE Based Training Operation

In order to enable the WMR to follow a predefined trajectory, it needs to get trained. The training set includes the reference trajectory and uncertainty range as input and actual trajectory as output. It involves two types of operations:

Forward operation: ANFIS trains itself using the gradient descent algorithm to optimize the original parameters, and then solve the resultant parameters using a least squares algorithm (Off-line Learning).

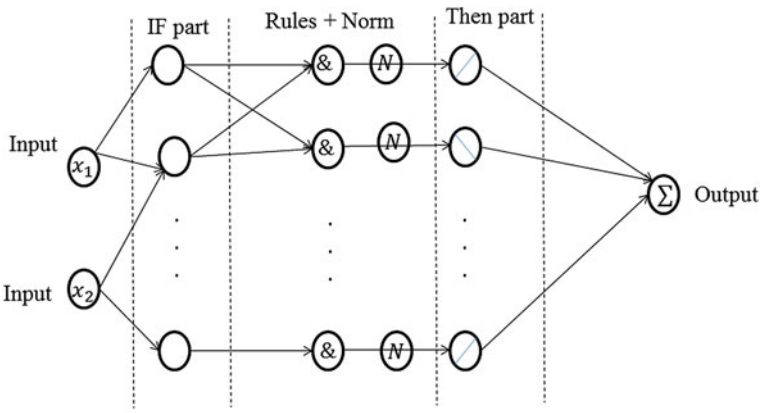


Fig. 9 Basic ANFIS network

Backward operation: After updating these parameters using a least squares algorithm, the former parameters are then updated by back-propagating the errors that are still present (Back propagation). The equation for parameter update by [9]:

$$\alpha_i(t+1) = \alpha_i(t) - \eta \frac{\partial E}{\partial \alpha_i} \tag{36}$$

$$E = (y - yd) \tag{37}$$

The ANFIS has the advantage that one may train it with the simplest available controller for example PID controller, LQR, LQI or SDRE controller. For the training of ANFIS an SDRE controller is being designed as follows:

State dependent Riccati equation (SDRE) is the nonlinear version of linear quadratic regulation. This method is based upon the method of extended linearization [10]. In the method of extended linearization the nonlinear dynamics of the system are considered constant at each particular time instant and then calculate the solution of Riccati equation. For this purpose consider the following general nonlinear system:

For this purpose consider the following general nonlinear system:

$$\dot{x} = f(x) + g(x)u$$

$$y = C(x). \tag{38}$$

The concept of extended linearization is to choose the dynamics at each instant in such a way that they are linear. The dynamics of such a system are expressed as:

$$\begin{aligned} \dot{x} &= A(x)x + B(x).u \\ y &= C(x)x. \end{aligned} \tag{39}$$

These dynamics $A(x)$, $B(x)$, $C(x)$ are not unique and the choice of these dynamics is called parameterization problem [11]. The performance of the controller depends on choice of the user. Therefore the SDRE concept is to

consider the system as a linear system at each instant and calculate the control law "on-line". The control law for tracking of WMR is given by [12] as:

$$u = -K(x)x \tag{40}$$

where $K(x)$ is given as:

$$K(x) = R^{-1}B(x)P(x) \tag{41}$$

And $P(x)$ is the solution of the state dependent Riccati equation:

$$A^T(x)P(x) + P(x)A(x) - P(x)B(x)R^{-1}B^T(x)P(x) + Q = 0. \tag{42}$$

This will minimize the cost function given as:

$$J = \int_0^t x^T(t)Qx(t) + u^T(t)Ru(t)dt \tag{43}$$

In the above equation $R(x)$ and $Q(x)$ is $m \times m$ and $n \times n$ weighting matrices respectively.

5.4 ANFIS Tracking Performance

In case of ANFIS the training set is obtained using an SDRE based trajectory tracking WMR, reference trajectory is taken as input and resultant trajectory is considered as the output. For training purposes, an ANFIS network is being generated using five nodes and a bell shaped Gaussian membership function. The square trajectory as being considered for training of ANFIS and then the results are checked for circular shaped trajectory. After this further simulations are performed by training ANFIS for the Different uncertainty cases. ANFIS is trained for the case of 10, 15 and 20 % uncertainty and then the performance is analyzed for 17 % uncertainty. The performance of ANFIS is further analyzed by keeping the uncertainty same that is using 10 % uncertainty (present in training data) and changing the reference trajectory at input and later on changing both the uncertainty and reference trajectory that is using 17 % uncertainty and circular trajectory at input (Figs. 10, 11, 12, 13, 14, 15, 16, 17).

6 Comparison of MRAC and ANFIS Controllers

This section is solely dedicated to the comparison of both the techniques in terms of complexity, speed and other specifications.

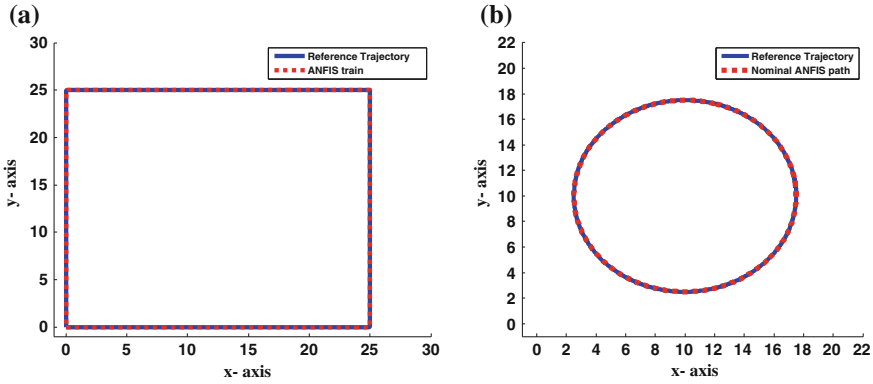


Fig. 10 ANFIS training results for nominal model. **a** Square trajectory. **b** Circular trajectory

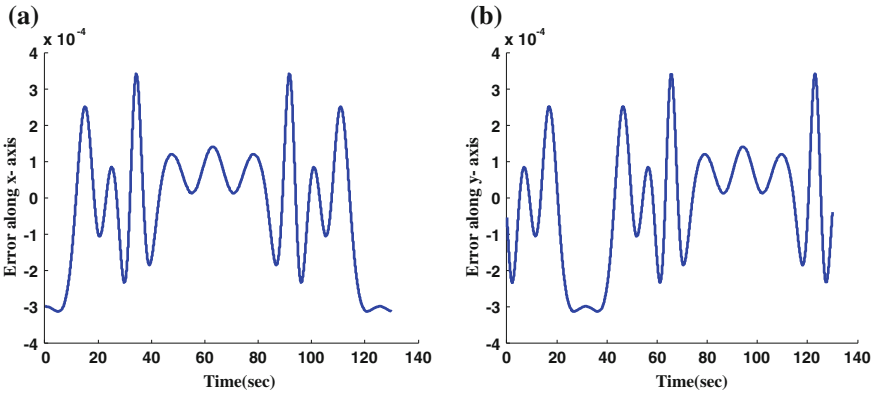


Fig. 11 Tracking error using ANFIS training. **a** x-axis. **b** y-axis

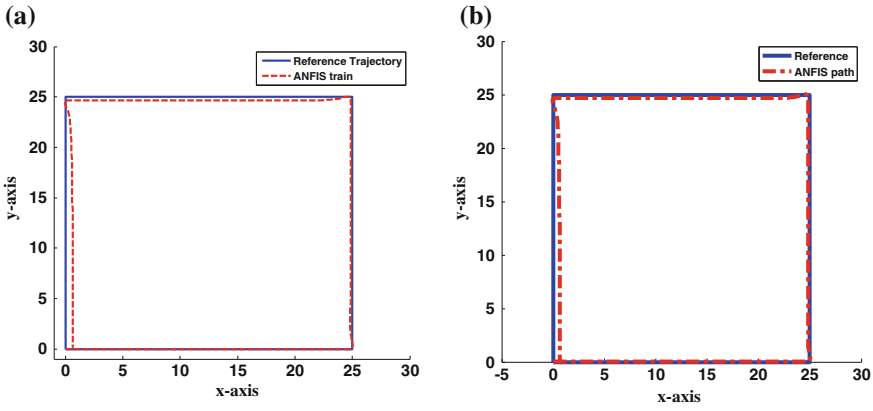


Fig. 12 Tracking error using ANFIS training. **a** 10 % uncertainty. **b** 17 % uncertainty

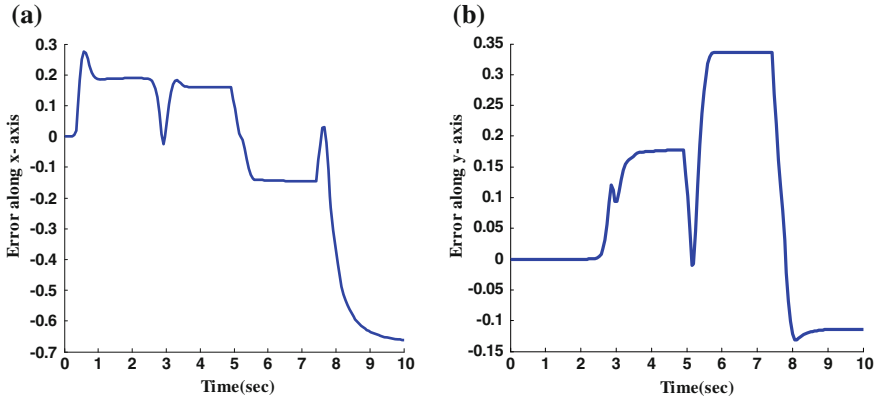


Fig. 13 Trajectory tracking error using ANFIS. **a** x-axis. **b** y-axis

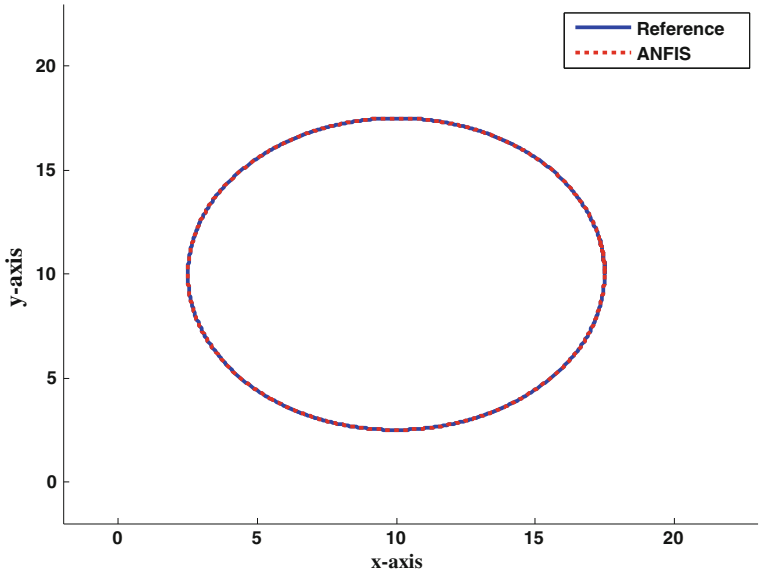


Fig. 14 ANFIS performance using different trajectory as input keeping 10 % uncertainty

6.1 Complexity

Although, ANFIS offers a complex network but it is very simple when compared to the tiresome derivation of MRAC controller. Moreover, the update law for MRAC controller is based on Lyapunov approach and the Lyapunov function itself is not unique for a particular problem. Thus, model free approach is more preferable when it comes to complexity.

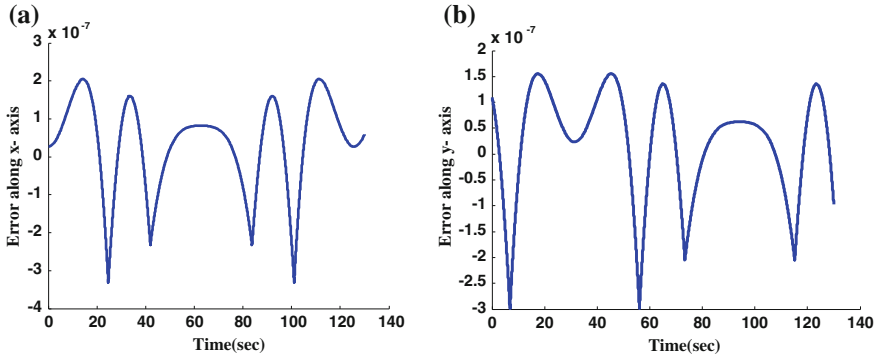


Fig. 15 Trajectory tracking error using ANFIS. **a** x-axis, **b** y-axis

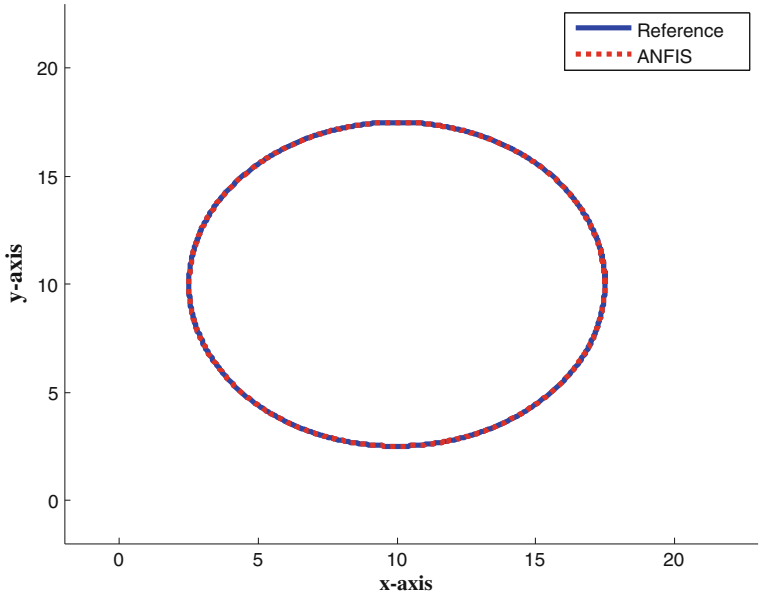


Fig. 16 ANFIS performance using different trajectory and 17 % uncertainty as input

6.2 Computational Time

It is clear from Sect. 5.1 that Model free approach offers less computational time than Model dependent methods.

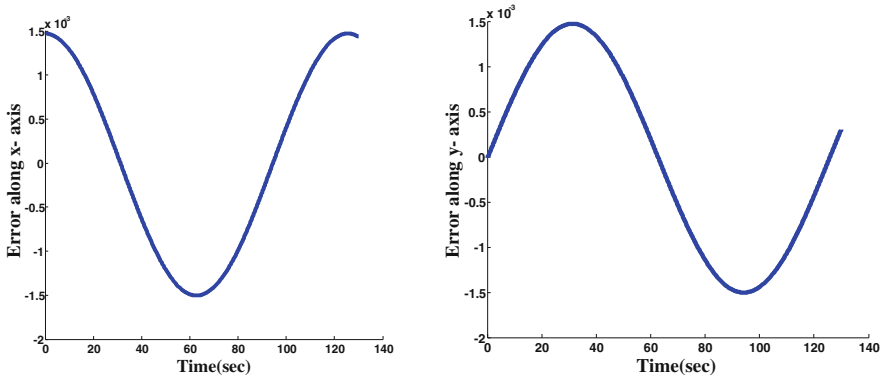


Fig. 17 Trajectory tracking error using ANFIS. **a** x-axis, **b** y-axis

6.3 Convergence Properties

In case of MRAC, the convergence of controller requires the parameters of WMR to be slowly time varying however, ANFIS faces no such constraints. Moreover, the convergence of ANFIS is much faster as compared to MRAC because of its lesser computational time.

6.4 Model-Dependence

MRAC works on the derived controller using the dynamic model of WMR therefore it is completely model dependent. Even though it is adaptive in nature but for update law to be derived basic model structure of WMR must be known. On the other hand, ANFIS is completely model free. IT only requires a training set in order to follow any desired trajectory.

6.5 Intelligence

Artificial intelligence in robots is in demand these days and it can be achieved using ANFIS. It just needs an initial example to learn about its task and then it can easily provide us with the required output. Although MRAC is adaptive to uncertainties and disturbances but it is certainly not intelligent as it requires a cumbersome task of designing a complete controller using WMR model.

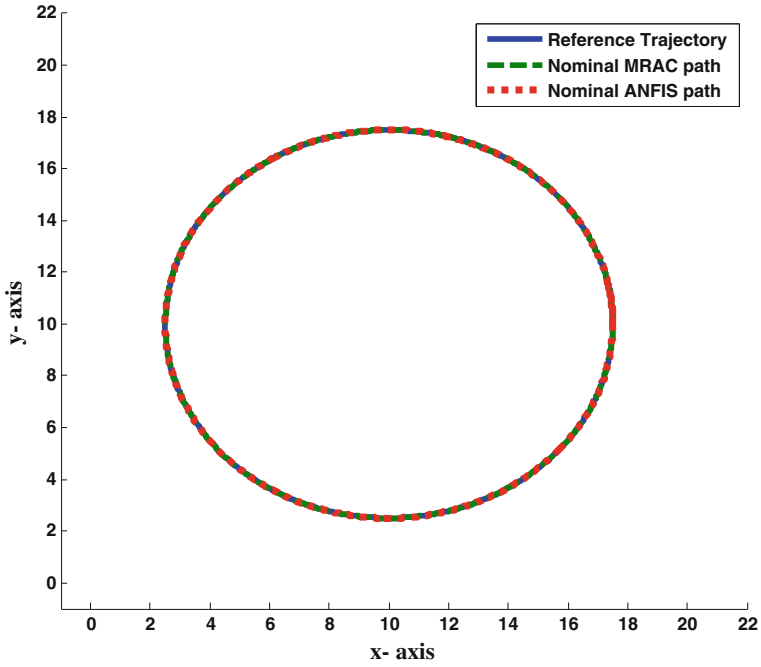


Fig. 18 Circular shaped trajectory tracking performance using MRAC and ANFIS (Nominal plant model)

6.6 Tracking Performance

In order to evaluate and compare the performance of both approaches simulations are performed using MATLAB/SIMULINK. The above mentioned tracking results are overlapped in order to compare the both approaches in terms of tracking performance (Figs. 18, 19, 20, 21, 22, 23, 24, 25).

Above discussion clearly signifies the importance and usefulness of Model free scheme in the coming technology era over model dependent methods. These types of artificially intelligent robots can meet almost any type of given desired path during industrial and navigational applications, thereby making them more and more useful. Following section presents a performance based comparison between both schemes.

7 Conclusion

This chapter presents two different schemes for the trajectory tracking of wheeled mobile robot. The first scheme is Model Reference adaptive controller (MRAC) which exemplifies the Model dependent approach and the second scheme represents

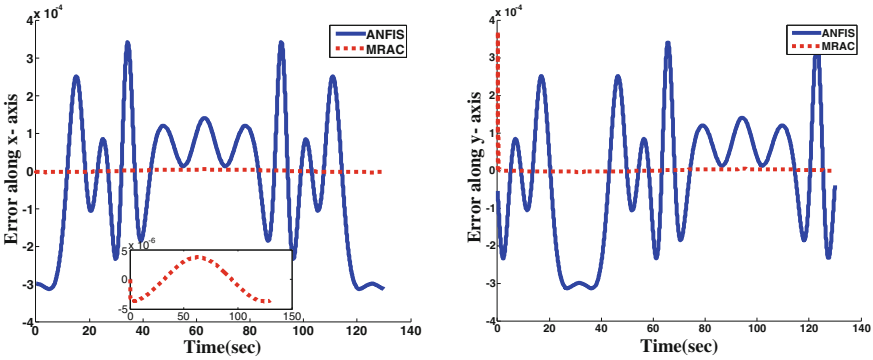


Fig. 19 Trajectory tracking error using MRAC and ANFIS. **a** x-axis, **b** y-axis

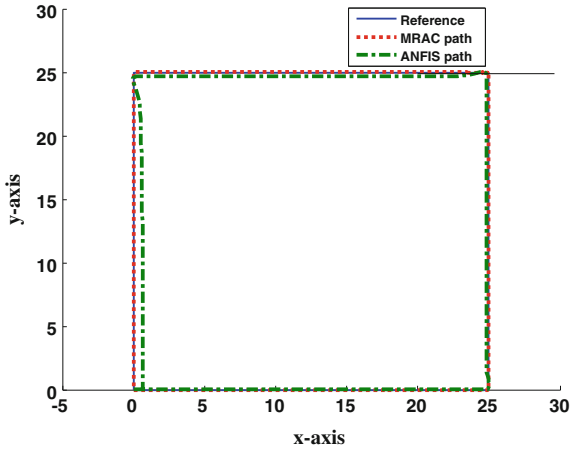


Fig. 20 Square trajectory tracking performance with MRAC and ANFIS (17 % uncertainty)

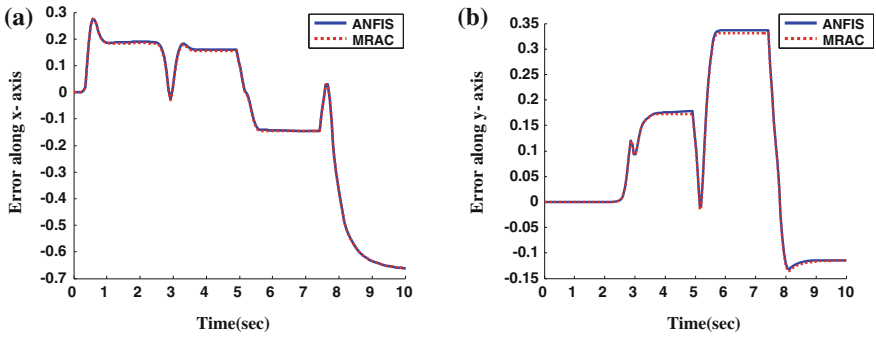


Fig. 21 Trajectory tracking error using MRAC and ANFIS. **a** x-axis, **b** y-axis

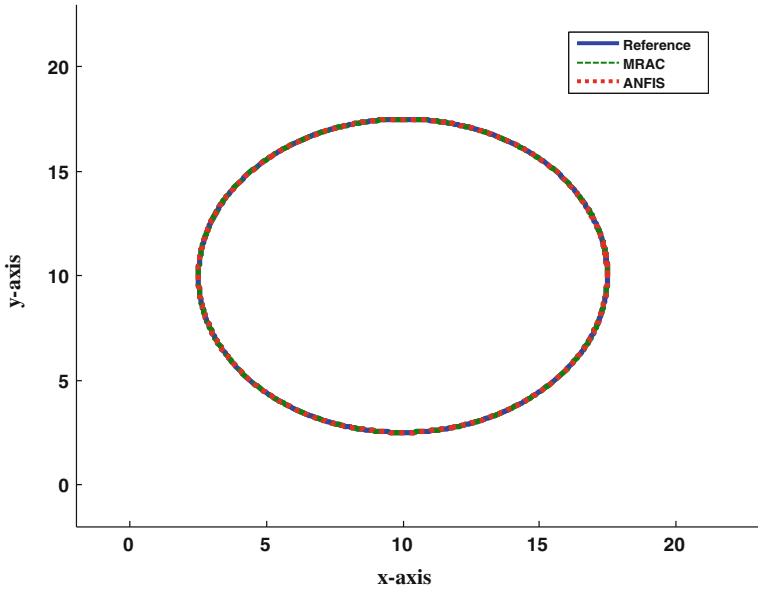


Fig. 22 Circular trajectory tracking performance with MRAC and ANFIS (10 % uncertainty)

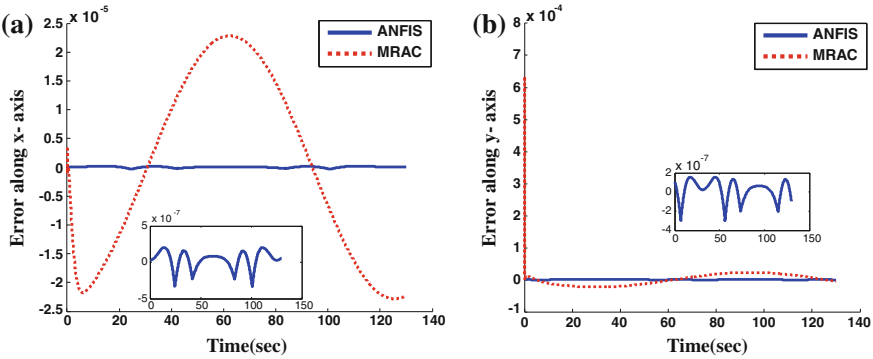


Fig. 23 Trajectory tracking error using MRAC and ANFIS. **a** x-axis, **b** y-axis

data driven approach which is Adaptive Neuro-Fuzzy Inference System (ANFIS). Both the techniques are explained and compared based upon the complexity, convergence, speed, model dependence and other criterion. All the principles used for comparison decided in the favor of ANFIS as a better approach. The simulation results also displayed a very minor error gap between the two approaches. Since the ANFIS was trained using a simple SDRE based controller which is also vulnerable to parametric uncertainty errors the simulations presented higher errors for the ANFIS case. However, the above simulations have proved that ANFIS is capable of

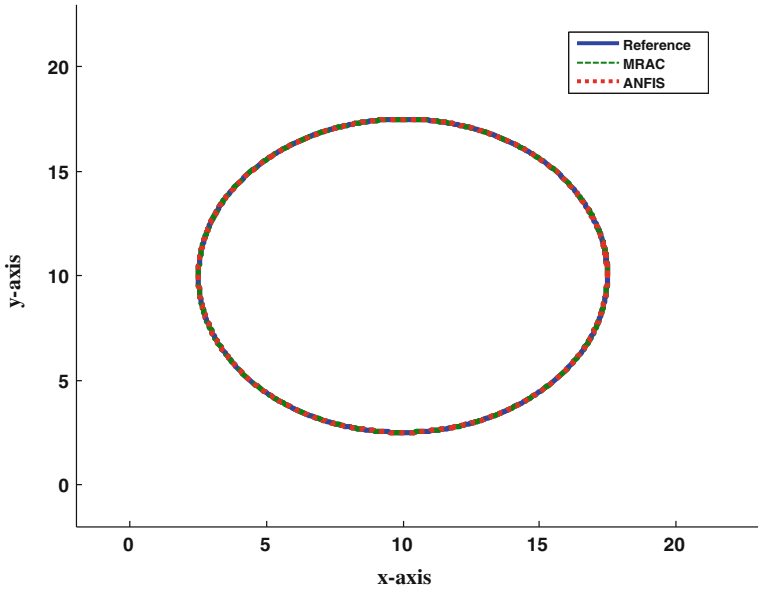


Fig. 24 Circular trajectory tracking performance using MRAC and ANFIS (17 % uncertainty)

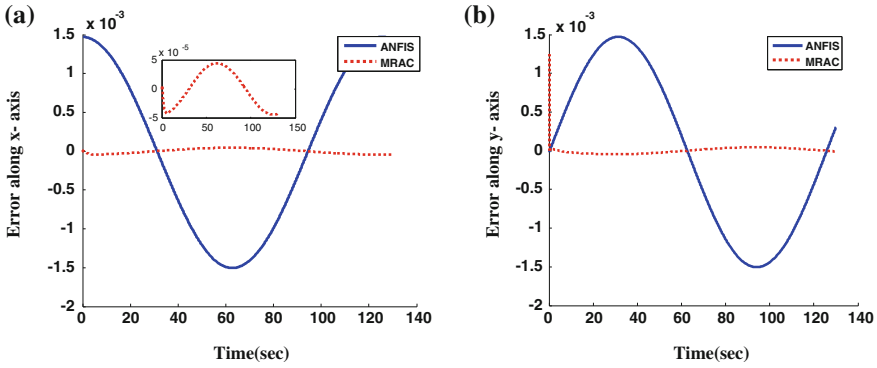


Fig. 25 Trajectory tracking error using MRAC and ANFIS. **a** x-axis, **b** y-axis

adapting in the presence of parametric uncertainties. Once trained on a particular trajectory using different uncertainty cases, it is capable of adapting to any trajectory and uncertainty within the limits. Although, the error of ANFIS is very slightly higher than MRAC in case of parametric uncertainty but if the simplicity and model-free nature of this approach is considered, it proves to be better than MRAC. Thus, one can simply state that model free approaches are much more appropriate solution for control applications as compared to the model dependent schemes when it comes to real world uncertain scenarios.

References

1. C. Stuart, *Encyclopedia of artificial intelligence*, 2nd edn. (Wiley, New York, 1992)
2. D. Ioan, L. Rogelio, M. Mohammad, K. Ali Reza, *Adaptive Control Algorithms, Analysis and Applications* (Springer, London, 2011)
3. C. Emine, O. Metin, Model reference adaptive control of wheeled mobile robot for trajectory tracking. Paper presented at IEEE conference on INISTA, 2012
4. J. John, H. Ping, Adaptive control of mechanical manipulators. Paper presented at IEEE international conference on robotics and automation, 1986
5. X. Xianbo, L. Lionel, L. Chao, J. Bruno, Path tracking: combined path following and trajectory tracking for autonomous underwater vehicles. Paper presented at IEEE international conference on intelligent robots and systems, USA, 2011
6. F. Pourboghrat, Adaptive learning control for robotics. Presented in IEEE international conference on robotics and automation, 1988
7. S. Khoshnam, M. Alireza, T. Ahmadreza, T. Behzad, Adaptive trajectory tracking control of a differential drive wheeled mobile robot. Robotica **29**, 391–402 (2011). doi:10.1017/S0263574710000202
8. N. Felipe, An adaptive dynamic controller for autonomous mobile robot trajectory tracking. Control Eng. Pract. **16**, 1354–1363 (2006)
9. T. Benmiloud, in *Multioutput Adaptive Neuro-fuzzy Inference System*, ed. by V Munteanu, R. Raducanu. Recent Advances in Neural Networks, Fuzzy Systems & Evolutionary Computing (Pearson Education Private Limited, Gurgaon, 2006)
10. F. Bernard, *Advanced Control System Design* (Prentice-Hall, Upper Saddle River, 1996)
11. J. Cloutier, N. Christopher , P. Curtis, Nonlinear regulation and nonlinear hinf control via the state-dependent riccati equation technique: part 2, examples. Paper presented at 1st international conference on the nonlinear problems in aviation and Aerospace, Daytona Beach, 1996
12. A. David, F. Bernard, State dependent differential Riccati equation for nonlinear estimation and control, in *IFAC, 15th Triennial World Congress*, Barcelona, 2002

Part IV
Power and Energy

Computational Intelligence in Smart Grids: Case Studies

Mohamed A. Abido, El-Sayed M. El-Alfy and Muhammad Sheraz

Abstract This chapter briefly provides an overview of related work on computational intelligence techniques in smart grids. It also reviews two computational intelligence techniques and some of their current applications in solving problems associated with smart grids implementation and deployment. More importantly, two case studies are presented and intensively discussed. These applications include parameter estimation of photovoltaic models and tracking of maximum power point. This chapter also highlights some open research problems and directions for future research work.

Keywords Computational intelligence · Differential evolution · Adaptive neuro-fuzzy inference system · Hybrid systems · Smart grids · Parameter estimation · Photovoltaic models · Maximum power point tracking

1 Introduction

The last decade has witnessed a phenomenal increase in the application of computational intelligence in the automation of power systems design, operation, and control. The term smart grids has emerged and become omnipresent. According to the European Technology Platform for Electricity Networks of the Future [1], a smart grid refers to "*an electricity network that can intelligently integrate the actions of all users connected to it—generators, consumers and those*

M. A. Abido · M. Sheraz
Electrical Engineering Department, College of Engineering, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia

E.-S. M. El-Alfy (✉)
College of Computer Sciences and Engineering, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia
e-mail: alfy@kfupm.edu.sa

Z. H. Khan et al. (eds.), *Computational Intelligence for Decision Support in Cyber-Physical Systems*, Studies in Computational Intelligence 540,
DOI: 10.1007/978-981-4585-36-1_9,

that do both—in order to efficiently deliver sustainable, economic and secure electricity supplies." This technology is promising and has many vital characteristics such as self-healing and reliable networks, bi-directional communication, real-time measurements, intelligent sensing, demand-side control, energy conservation, asset management, etc. [2]. Efficient use of the limited resources that are available can be achieved through optimal solutions during power generation, transmission and distribution. Other goals include constructing and operating power systems that can provide affordable, reliable, secure and sustainable global supply of energy under normal and abnormal operating conditions [3–5].

Due to the complexity and challenges in the design, optimization, scheduling and management of smart grids, computational intelligence techniques have attracted much attention of the multi-disciplinary research community in recent years [6–13]. Many of the existing and emergent problems in smart grids have been addressed using techniques such as neural networks, evolutionary computation paradigms, fuzzy logic, swarm intelligence, immune systems, self-organizing maps and hybrid approaches [14–17]. Such powerful techniques allow building adaptive and optimal solutions that can cope with nonlinearities, changes and uncertainties in such complex large-scale environments. Jiang [18] identified several potential areas of application of computational intelligence techniques in smart electric grids. Venayagamoorthy [19] briefly described some potentials and promises of computational intelligence techniques for realizing intelligent smart grids. Some other specific applications are described in [15] using genetic algorithms, particle swarm and recurrent neural network.

Among the important aspects of smart grids are intelligence, decentralization and autonomy of control mechanisms. Carvalho et al. [20] presented a multi-agent based power management system using reinforcement learning for smart grid environments. They assumed renewable power sources with unpredictable variations in the output which can be caused by transient effects of weather or cloud. The software agents track and report the variations in the power flow at different locations in the grid. They also share information and control actions to coordinate the configurable energy storage devices to maintain the energy requirements to consumers. The agents learn from experience to improve their control actions to efficiently deal with different conditions. Another application of reinforcement learning for smart electric grid environments is described in [21]. The authors proposed a class of autonomous broker agents that learn to optimally derive long-term maximum profit policies for retail electricity trading using arbitrary economic signals from their environments. Two relatively recent reviews of the evolution of smart grid concepts with examples of computational intelligence in meeting their needs are presented in [5, 17].

This chapter briefly provides an overview of related work on computational intelligence techniques in smart grids. It also reviews two computational intelligence techniques and some of their current applications in solving problems associated with smart grids implementation and deployment (Sect. 2). More importantly, two case studies are presented and intensively discussed (Sect. 3). These applications include parameter estimation of photovoltaic models and

tracking of maximum power point. This chapter also highlights some open research problems and directions for future research work (Sect. 4).

2 Methods

In this section, we review two computational intelligence (CI) methods. The first one is Differential Evolution (DE), which is an efficient Evolutionary Algorithm (EA). The second method is Adaptive Neuro-Fuzzy Inference System which a hybrid computational intelligence technique. In the following sections these methods will be utilized for parameter estimation in photovoltaic systems and for maximum power point tracking.

2.1 Differential Evolution

Differential evolution (DE) is a population-based optimization technique which was first introduced in [22]. It is renowned for its robustness, simplicity, rapid convergence and less control variables. Moreover, DE is suitable for non-differentiable and nonlinear optimization with the ability to search for global optimum regardless of the initial values of the parameters. Like other evolutionary algorithms and search techniques, DE needs to maintain a population of a number of candidate solutions; usually called parameter vectors, individuals or chromosomes. Depending upon the dimension of the problem to be solved, each individual encodes the control variables or the optimizing parameters. Let G_i refer to the population of candidate solutions; it can be represented as follows:

$$G_i = [X_1, X_2, \ldots, X_{NP}] \tag{1}$$

where i is an index referring to the generation (iteration) number, X_j refers to the j-th individual, NP refers to the number of individuals or population size. Each individual is represented by an n-dimensional vector of the control variables x_k:

$$X_j = [x_1, x_2, \ldots, x_D] \tag{2}$$

where D is the dimension of the problem.

In all search algorithms the key step is to form the trail vector (variant vector) from the parent vector. The stratagem employed by DE to generate a trail vector is based on the difference between randomly chosen individuals. A trail (new) individual is generated for every parent (old) individual using the operations of mutation and crossover. The best individual is selected for the next generation (iteration) by comparing the objective function of old and new individuals. DE optimization process resembles the genetic algorithm (GA) with little difference.

While GA utilizes a crossover operation as a search method, DE employs a mutation operation as a search mechanism. DE includes following steps:

Step 1: Initialization
As a first step, DE needs to initialize its following control parameters:

G_{max}:	Number of generations (iterations), usually used as a stopping criterion
NP:	Population size (number of candidate solutions)
D:	Problem dimension or number of control variables (optimizing parameters)
$x_{k,\min}$ and $x_{k,\max}$:	Lower and upper bounds of the k-th control variable
F:	Mutation factor, where $0 \leq F \leq 1$
CR:	Crossover rate, where $0 \leq CR \leq 1$.

It starts the optimization process by generating an initial population containing the NP candidate solutions, each having D control variables. The values of the variables are selected randomly to lie within their minimum and maximum limits by following equation;

$$x_k = x_{k,\min} + rand \times \left(x_{k,\max} - x_{k,\min}\right) \tag{3}$$

where *rand* is a random number between 0 and 1.

Step 2: Locating the best solution
After initialization, the objective function is evaluated for each individual and the best solution (individual) is determined. The best solution is then updated by comparing is value with the next generation best solution to locate the global optimum.

Step 3: Mutation
Mutation means a perturbation with a random element. It is a first operation that is utilized to generate a trail (variant) vector and creates a "mutant vector" for all individuals in the current generation. For every individual of parent vector, X_i^g, in the current generation, g, two individuals $\left(X_a^g, X_b^g\right)$ are randomly selected from the population. A mutant vector, V_i^g, is then generated by adding a weighted difference of a best vector, X_{best}^g, and the parent vector, X_i^g, and a weighted difference of two randomly selected vectors $\left(X_a^g, X_b^g\right)$ with a base individual using the following equation:

$$V_i^g = X_i^g + F \times \left(X_{best}^g - X_i^g\right) + F \times \left(X_a^g - X_b^g\right) \tag{4}$$

where F represents the mutation factor having real value in the range [0, 1]; its value affects the speed of convergence.

Step 4: Crossover
To further improve the diversity and add more perturbation, a crossover operation is applied. In this operation, the mutant vector, V_i^g, and its parent vector, X_i^g, are combined to form a trial solution, U_i^g. The control variables of both vectors are

mixed in this step based on the crossover factor (*CR*) to form the trail solution. The crossover process can be specified by the following equation:

$$u_{i,j}^{g} = \begin{cases} v_{i,j}^{g} & rand \leq CR \\ x_{i,j}^{g} & otherwise \end{cases} \tag{5}$$

where *CR* is the crossover factor having a value in the range [0, 1], $1 \leq i \leq NP$ and $1 \leq j \leq D$.

Step 5: Selection
The selection operation is considered to be the last step in the formation of a new population. In this stage, the objective function value of the generated trail vector is reckoned and comparison is made with the corresponding value of the parent vector. If the individuals of the trail vector have better objective value than the corresponding individuals of parent vector, they will replace them; otherwise the parent vector will be retained. The selection operation for the minimizing problem can be given by:

$$X_i^{g+1} = \begin{cases} U_i^{g} & J(U_i^{g}) \leq J(X_i^{g}) \\ X_i^{g} & otherwise \end{cases} \tag{6}$$

where $J(.)$ is the objective function.

Step 6: Stopping Criteria
As DE is an iterative-based technique, it requires certain stopping criteria to end the looping process. They can be any user-defined conditions. Usually a predefined value of the maximum number of generations (iterations), a pre-specified tolerance in the error, or a combination of both is used for that purpose. In this chapter, the stopping criteria are based on objective function values and the algorithm stops when the best objective function value remains the same for a prescribed number of generations.

2.2 Adaptive Neuro-Fuzzy Inference System

The Adaptive Neuro-Fuzzy Inference System (ANFIS) [23] is a hybrid approach composed of an artificial neural network and a fuzzy inference system. It has a network-type structure and maps the input–output data set using the parameters of fuzzy membership functions. Figure 1 demonstrates a typical ANFIS architecture based on a two-rule Sugeno system with two inputs (x and y) and a single output, F. Assume (A_1, A_2) and (B_1, B_2) are the fuzzy input memberships for the inputs x and y, respectively, and they are used to fuzzify the inputs.

A two-rule Sugeno ANFIS has rules of the form:

$$IF\ x \text{ is } A_1 \text{ AND } y \text{ is } B_1 \text{ THEN } f_1 = p_1 x + q_1 y + r_1 \tag{7}$$

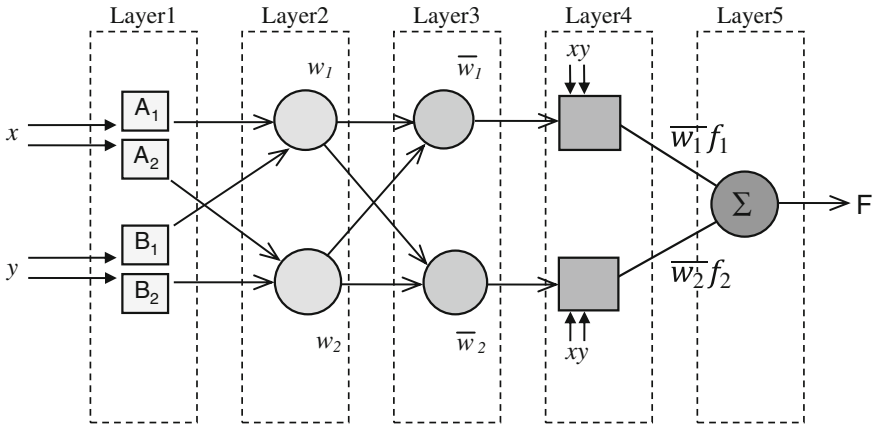


Fig. 1 A typical ANFIS architecture

$$IF\ x \text{ is } A_2 \text{ AND } y \text{ is } B_2 \text{ THEN } f_2 = p_2x + q_2y + r_2 \tag{8}$$

The overall ANFIS architecture has five layers and now we will discuss each of them in turn.

Layer 1

In this layer, every node is adaptive node and their number depends upon the number of input membership functions. Their output is given by:

$$O_{1,i} = \mu_{A_i}(x); \quad \text{for } i = 1, 2 \tag{9}$$

$$O_{1,i} = \mu_{B_{i-2}}(y); \quad \text{for } i = 3, 4 \tag{10}$$

where the subscripts 1 and i represent the layer number and node number, respectively; $\mu_A(.)$ and $\mu_B(.)$ are the membership functions; and $O_{1,i}$, for $i = 1$ to 4, are the membership values corresponding to the crisp inputs x and y.

The membership functions $\mu_A(.)$ and $\mu_B(.)$ can be any shaped functions like trapezoidal, triangle, or Gaussian. The most commonly used membership function $\mu(.)$ is the generalized bell shaped function which is given by:

$$\mu(x) = \frac{1}{1 + \left|\frac{x - c_i}{a_i}\right|^{2b_i}} \tag{11}$$

where a_i, b_i and c_i are parameters of the membership function (called premise parameters) and need to be optimized during the training process.

Layer 2

Every node in this layer is a fixed node and accepts the output (membership values) from layer 1 where t-norm is utilized to "AND" these values, as given by:

$$O_{2,i} = w_i = \mu_{A_i}(x)\mu_{B_i}(y); \text{ for } i = 1, 2 \tag{12}$$

The output of each node corresponds to the firing strength of a rule.

Layer 3
Every node in this layer is a fixed node and is used to normalize the firing strength by dividing the rule's firing strength by the sum of all rules firing strengths, as given by:

$$O_{3,i} = \bar{w}_i = \frac{w_i}{w_1 + w_2}; \text{ for } i = 1, 2 \tag{13}$$

The output of each node represents the normalized firing strength of a rule.

Layer 4
Every node in this layer is an adaptive node and is given by the function:

$$O_{4,i} = \bar{w}_i f_i = \bar{w}_i(p_i x + q_i y + r_i) \tag{14}$$

where p_i, q_i and r_i are the consequent parameters and need to optimized in the training process.

Layer 5
It has only one fixed node which sums up all the input signals to get the final output as given by:

$$O_{5,i} = \sum_i \bar{w}_i f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \tag{15}$$

Learning Process
In the learning algorithm, ANFIS optimizes and adapts its parameters using the training data set to predict the output data with a high accuracy. The Sugeno-type model has two types of parameters [24]:

- Nonlinear parameters or membership functions' parameters (premise parameters).
- Linear parameters or rules' parameters (consequent parameters).

Various learning methods have been proposed by researchers. The method used in this paper is based on the hybrid learning algorithm that employs a combination of back-propagation (BP) and least-square estimation (LSE) to optimize the premise and consequent parameters. This method has two passes: forward pass and backward pass. In the forward pass, the consequent (linear) parameters are calculated using an LSE algorithm while the premise (nonlinear) parameters are unmodified. In the backward pass, the premise (nonlinear) parameters are calculated using a back-propagation algorithm while the consequent (linear) parameters are unmodified.

The LSE learning algorithm calculates the square error between the actual output and the predicted output obtained from the Sugeno-type model. This error is

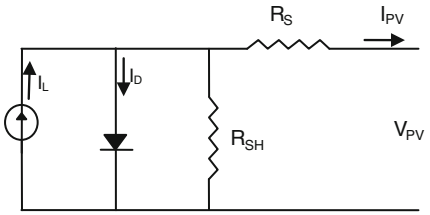


Fig. 2 Equivalent electric circuit model of a PV device

utilized to adapt the consequent parameters of the model. The back-propagation gradient descent method uses the error between the actual output and predicted output in the backward pass to calculate the error in different nodes.

3 Case Studies

3.1 Case Study 1: Parameter Estimation of Photovoltaic Models

The electrical characteristics of a photovoltaic (PV) can be represented by an equivalent electric circuit model as shown in Fig. 2. It consists of light-dependent current source, a p-n junction diode and two resistances (one in series and another in parallel). This model is widely known as five-parameter equivalent electric circuit model (since it has five unknown parameters). These parameters are the light generated current (I_L), reverse saturation current (I_o), series resistance (R_S), shunt resistance (R_{SH}) and modified ideality factor (a). The performance and competence of this model is highly dependent on the values of these parameters and must be known for its implementation.

By Kirchhoff's law, the simple relationship of currents can be found as given by:

$$I_{PV} = I_L - I_0\left\{\exp\left[\frac{V_{PV} + I_{PV}R_S}{a}\right] - 1\right\} - \left[\frac{V_{PV} + I_{PV}R_S}{R_{SH}}\right] \tag{16}$$

$$a = \frac{N_s nkT}{q} \tag{17}$$

where I_{PV} and V_{PV} are the PV panel current and voltage, respectively; N_S represents the number of cells in the PV panel; T is the cell temperature; n is the ideality factor of the diode; the constants k and q are Boltzmann's constant and electronic charge, respectively. Equation (16) represents the I–V characteristics of the PV and is governed by the values of the five parameters (I_L, I_0, R_S, R_{SH}, a). The modeling of a PV device is quite complex because of the transcendental nonlinear characteristics of Eq. (16) and the unknown values of the five parameters. Having known values of these parameters, Eq. (16) can be solved using an efficient

numerical method such as Newton–Raphson's method. With different atmospheric conditions, these parameters have different values which can be calculated at any ambient condition using the following translational model equations (Eqs. 18–23):

$$a = a_{ref}\left(\frac{T_C}{T_{C,ref}}\right) \tag{18}$$

$$I_L = \frac{S}{S_{ref}}\left[I_{L,ref} + \mu_{I,SC}\left(T_C - T_{C,ref}\right)\right] \tag{19}$$

$$R_{SH} = R_{SH,ref}\left(\frac{S_{ref}}{S}\right) \tag{20}$$

$$R_S = R_{s,ref} \tag{21}$$

$$\frac{I_0}{I_{0,ref}} = \left(\frac{T_C}{T_{C,ref}}\right)^3 \exp\left[\left(\frac{N_S T_{ref}}{a_{ref}}\right) * \left(\frac{E_{g,ref}}{T_{ref}}\left| - \frac{E_g}{T}\right|\right)\right] \tag{22}$$

$$\frac{E_g}{E_{g,ref}} = 1 - C\left(T - T_{ref}\right) \tag{23}$$

where S and T_c represent the solar radiation and temperature of the PV panel, respectively; $\mu_{I,sc}$ and N_S are the coefficient of the short-circuit current and the number of cells in the panel, respectively (both of these quantities are provided by the manufacturer); Eg is the band-gap energy of the PV cell material; $C = 0.0003174$. The quantities with the subscript *ref* represent their values at the Standard Test Condition (STC) (1000 W/m^2 and 25 °C).

The major challenge in the implementation of this model lies in the estimation of the model parameters as their values are not provided in the solar panel catalogue. The exact values of these parameters are required to regenerate the output characteristics of the PV panel accurately. Various algorithms have been presented in the literature; some of them require experimental data which is not usually provided by the manufacturer [25]. Other algorithms make few assumptions which deteriorate the performance of the model [26]. The efficiency of these later methods is tainted because of the assumptions made. Recently, several evolutionary algorithms, like Genetic algorithms (GA) [27] and particle-swarm optimization (PSO) [28], have been utilized to optimize these model parameters with promising results shown. In the following subsections, we describe an alternative method for estimating the model parameters using differential evolution (which was reviewed in Sect. 2.1).

3.1.1 Parameter Estimation Using DE

Here, we apply the DE-based parameter estimation method to determine the optimal values for the parameters involved in the photovoltaic model. DE requires

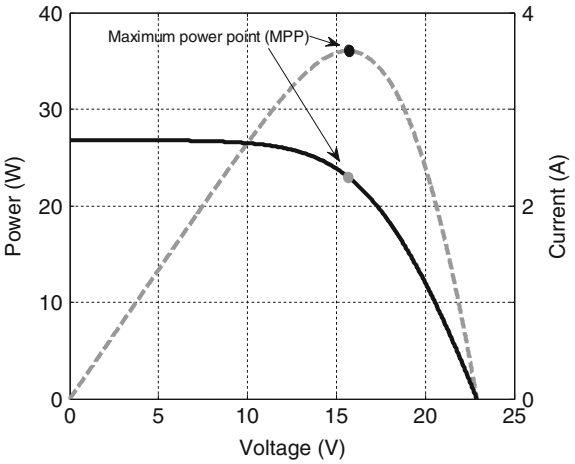


Fig. 3 Typical I–V and P–V curve of PV

only the data supplied in the manufacturer provided catalogues. This data gives the values of I–V pair at short-circuit condition (SC), open-circuit conditions (OC) and maximum power conditions (MP) at STC (1,000 W/m^2 and 25 °C). Figure 3 shows these points on the typical I–V and P–V curves. This method uses only the values of these points in the estimation process and can redraw high precision I–V and P–V curves of the PV.

Similar to other optimization techniques, DE requires an objective function that needs to be minimized or maximized. An appropriate objective function is needed for this purpose and should be dependent on the optimizing parameters (I_L, I_o, R_S, R_{SH}, a). The objective function used in this work is based on the error calculation and is given by the following equations:

$$J = Error_{MPP}^{V,I,P} + Error_{OC\text{ and }SC}^{V,I} \tag{24}$$

$$Error_X = \frac{X_{Meas} - X_{Exp}}{X_{Exp}} \tag{25}$$

where V_{MP}, I_{MP} and P_{MP} represent the voltage, current and power at maximum power point, respectively; V_{OC} and I_{SC} are the voltage at open circuit point and the current at short circuit point, respectively. The values of these points are given in the PV panel datasheet. V_{SC} and I_{OC} are the voltage at the short-circuit condition and the current at the open-circuit condition, respectively, having values equal to zero. X represents any of the above mentioned points. X_{Meas} is calculated using the estimated parameters and X_{Exp} is taken from the datasheet. *Error* represents the deviation of the measured value from the experimental value given in the datasheet.

The goal of this optimization problem is to identify the optimum values of the five parameters (I_L, I_0, R_S, R_{SH}, a) at STC by minimizing the objective function given in Eq. (24), which will basically reduce the error at the above mentioned key points.

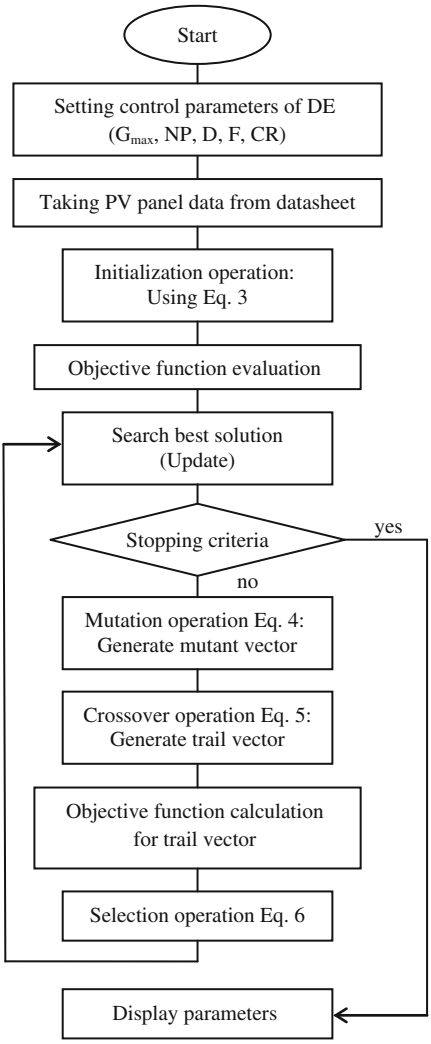


Fig. 4 The flowchart for the parameter estimation method using DE

The parameters determined at STC can then be used to calculate the values at other operating conditions using Eqs. (18–23) as stated before. The flowchart for the proposed parameter estimation method using DE is shown in Fig. 4.

3.1.2 Results and Discussion

The efficiency of the DE-based parameter estimation method is validated by comparing the determined curves (generated by estimated parameters) with the experimental curves. The experimental curves data is extracted from the PV panel

Table 1 Specifications of selected PV panels (taken from the datasheet)

Parameter	Mono-crystalline (m-Si) (Sun power 230 W)	Thin film (CIS) shell ST36
Open-circuit voltage (V_{OC})	48.7 V	22.9 V
Short-circuit current (I_{SC})	5.99 A	2.68 A
Maximum power voltage (V_{MP})	41 V	15.8 V
Maximum power current (I_{MP})	5.61 A	2.28 A
Maximum power (P_{MP})	230 W	36 W
Number of cells in series (N_S)	72	42
I_{SC} temperature coefficient (μ-I_{SC})	3.5 mA/°C	0.32 mA/°C
Voc temperature coefficient (μ-V_{OC})	–132.5 mV/°C	–100 mV/°C

Table 2 Control parameters of DE

Control parameters	Value used
Number of generation (G_{max})	200
Population size (*NP*)	250
Crossover factor (*CR*)	0.6
Mutation factor (*F*)	0.4

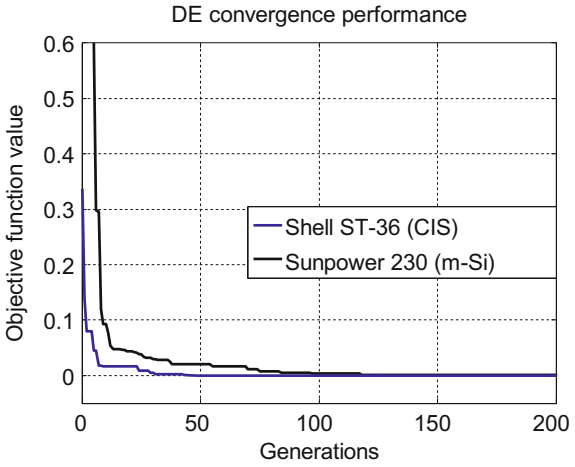


Fig. 5 Convergence performance of DE

datasheets using digitizer software. Two PV panels of different technologies are used for this purpose; these are mono-crystalline (m-Si) and thin film (CIS) and their specifications are given in Table 1.

The control parameters of DE are shown in Table 2. These values are selected by running the simulation several times and checking its convergence efficiency. Figure 5 shows the convergence performance of DE for the two selected PV panels. As shown, the values of the objective function become zero after 50 and 100 iterations for CIS and m-Si solar panel, respectively. The optimized values of the five parameters are shown in Table 3.

Table 3 Values of the five parameters identified using DE

Parameter	Mono-crystalline (m-Si) (Sun power 230 W)	Thin film (CIS) shell ST36
Light current "I_L" (A)	5.99	2.6803
Saturation current "I_0" (A)	1.4032e − 7	4.11965e − 05
Series resistance "R_S" (Ω)	0.008686	1.3901
Parallel resistance "R_{SH}" (Ω)	95658.60	38544.6
Modified ideality factor "a"	2.7715	2.0662

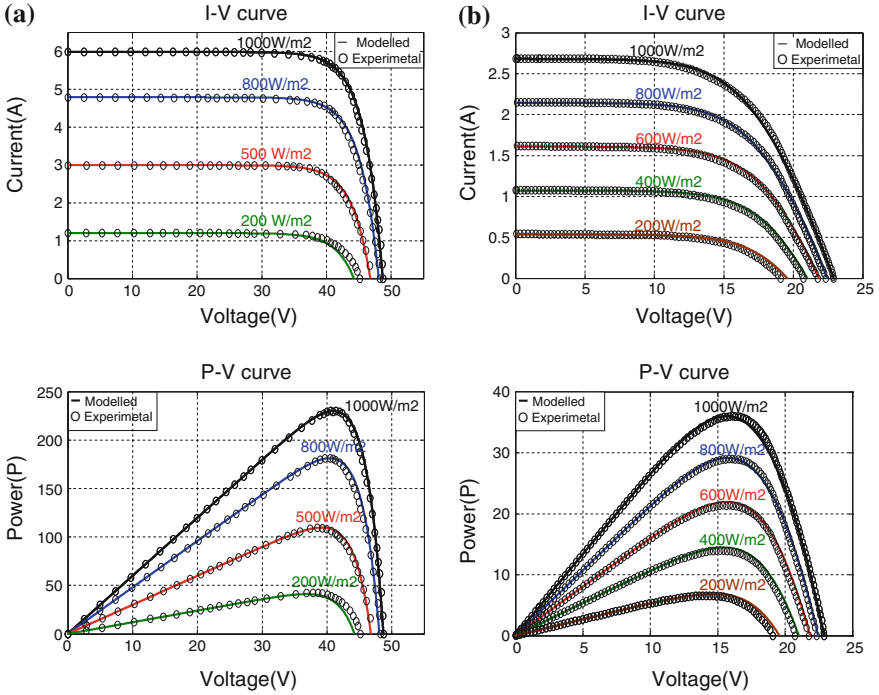


Fig. 6 The I–V and P–V curves for different irradiation levels and constant temperature 25 °C: **a** mono-crystalline (m-Si), **b** thin film (CIS)

3.1.3 Comparison with Experimental Data

Figure 6a, b shows the I–V and P–V curves generated from the estimated parameters for mono-crystalline (m-Si) and thin film (CIS) panel, respectively, for different irradiation levels and constant temperature of 25 °C. These figures illustrate that the I–V curves obtained from the proposed method are in great accordance with the experimental curves for all the irradiation levels, particularly

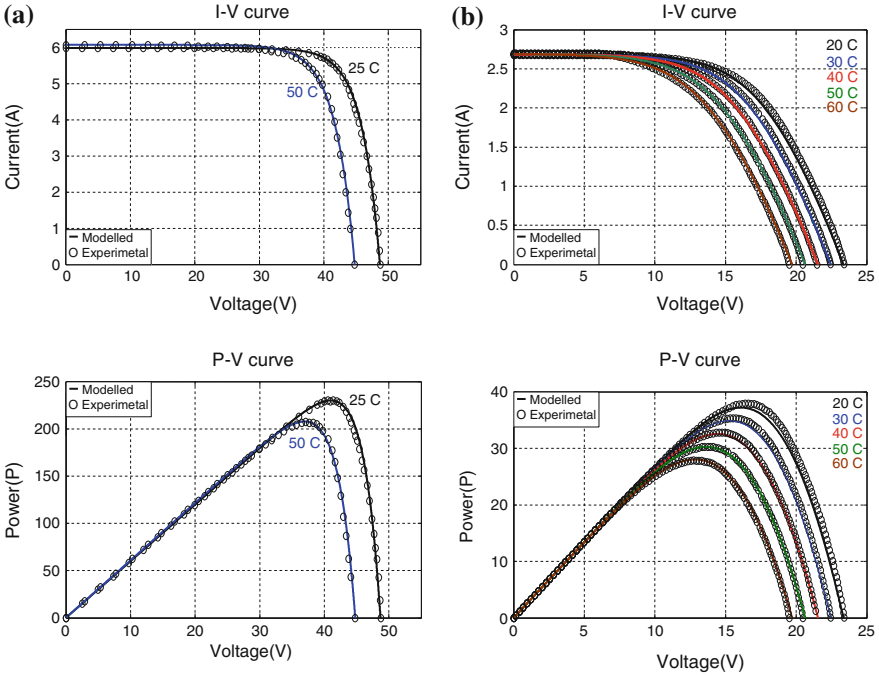


Fig. 7 The I–V and P–V curves for different temperature changes: **a** mono-crystalline (m-Si), **b** thin film (CIS)

for STC. The accuracy of the proposed method encountered with temperature changes is shown in Fig. 7 for both the technologies. The performance of the proposed approach is examined at 20 °C, 30 °C, 40 °C, 50 °C and 60 °C for thin film. For mono-crystalline, it is evaluated at two levels 25 °C and 50 °C because the experimental data for these temperatures only is available in the datasheet. It can be seen that the curves generated from the proposed method matches with the experimental curves under all the temperature variations.

3.1.4 Statistical Analysis

Three statistical errors are used for a comprehensive comparison of proposed approach. These errors are the root-mean-square error (RMSE), the mean-bias error (MBE) and the mean-absolute error (MAE). The relative values of these errors are given by the following equations:

$$RMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2}}{\frac{1}{n}\sum_{i=1}^{n} x_i} \tag{26}$$

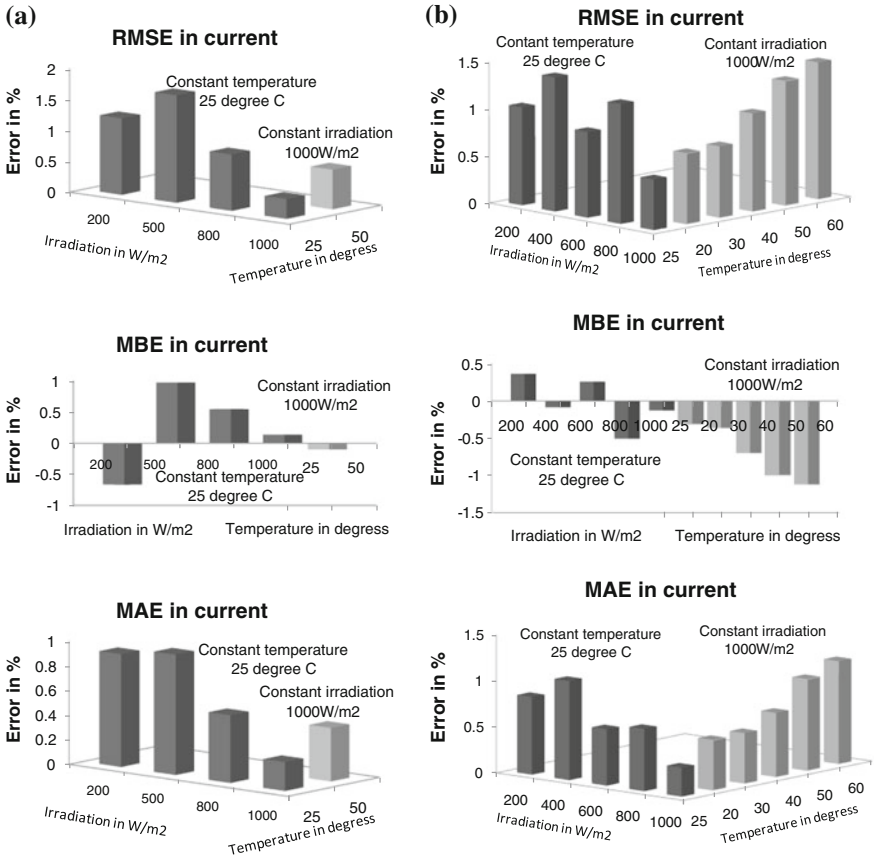


Fig. 8 The RMSE, MBE, MAE for different irradiations and temperatures: **a** mono-crystalline (m-Si) and **b** thin film (CIS)

$$MBE = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)}{\frac{1}{n}\sum_{i=1}^{n} x_i} \tag{27}$$

$$MAE = \frac{\frac{1}{n}\sum_{i=1}^{n}|y_i - x_i|}{\frac{1}{n}\sum_{i=1}^{n} x_i} \tag{28}$$

where y_i and x_i represent the measured and experimental values, respectively and n is the number of data points. The precision of the proposed method is analyzed by RMSE and MAE while MBE is used to show whether the approach is over-predicting or under-predicting the experimental values. These errors are calculated for current values computed against the voltage points. Figure 8 illustrates the values of these errors for different irradiation and temperature variations. It can be

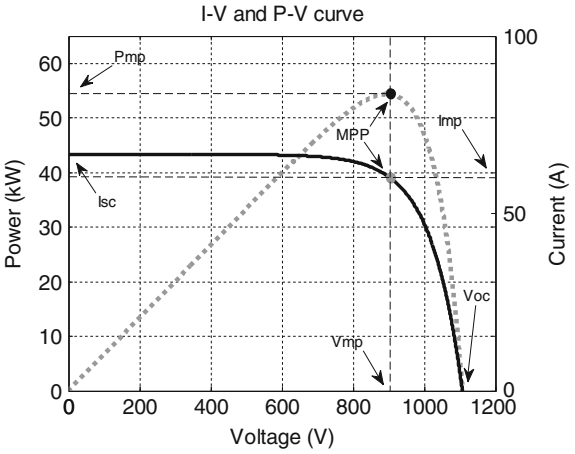


Fig. 9 The I–V and P–V curves of a PV device showing key points

observed that the error at STC is negligible and within the acceptable range for other conditions.

3.2 Case Study 2: Maximum Power Point Tracking

The output characteristics of a PV device are represented by I–V and P–V curves as shown in Fig. 9. For an arbitrary PV array configuration and a set of operating conditions, it is possible to plot an I–V curve showing the array's output current as a function of its output voltage. The array's P–V curve shows the output power as a function of the output voltage. Figure 9 shows two such curves for an arbitrary operating condition; the solid line shows the I–V curve of a PV array while the dotted line shows the power output of the same array. The small dot marked on the P–V curve is known as the Maximum Power Point (MPP) and it represents the maximum power that can be delivered by the PV array under a specific set of operating conditions. This is the point where the PV device is most efficient in converting the solar energy into electrical energy. The MPP is not a fixed point but actually varies throughout the day depending upon the environmental conditions, i.e. solar radiation and cell temperature. These conditions are commonly known as operating conditions for the PV device and are always changing with time which keeps varying the MPP. Therefore, the maximum power point tracking (MPPT) controller is of a great importance and an integral part of all kinds of the PV systems [36]. It forces the PV system to operate at its maximum efficiency.

Despite the increasing use of PV, this technology still faces a major obstacle due to its high capital cost and low efficiency. The overall efficiency of the PV system depends upon the efficiencies of PV panels, power electronics converters and maximum power point algorithm. The PV panels have efficiency around 8–20 % only, converters have 95–98 % and MPPT algorithm has more than 98 %.

The efficiencies of electronic converters and PV arrays depend on the technology but MPPT efficiency can be increased by improving its tracking methods. These efficient and superior techniques can easily be installed in the previously existed PV systems by updating their software with less cost.

Many MPPT techniques have been proposed in the literature and comprehensive comparisons of these methods have been presented in [29]. The conventional MPPT techniques are based on the approximation [30] and hill climbing methods [31] and have some deficiency such as oscillation in the steady state, large convergence time and possible failure to track MPP under the rapidly changing environmental conditions [32]. In [33, 34], the researchers improved these conventional methods by making some modifications but are not able to surmount these deficiencies completely. Recently, some intelligent techniques [16, 35] are used to overcome the shortcomings of the conventional methods and improve the tracking results.

3.2.1 ANFIS-Based MPPT Controller

As the output characteristics of a PV system are highly nonlinear, computational intelligence techniques are widely used to improve the efficiency of the MPPT controller. Fuzzy logic can transform the linguistic and heuristic terms to numerical values and vice versa using membership functions and fuzzy rules. Artificial neural network (ANN) can approximate the nonlinear functions to map inputs and outputs but it do not have heuristic nature. Researchers combine the fuzzy inference system (FIS) with the ANN to build a hybrid system named as Adaptive Neuro-Fuzzy Inference System (ANFIS) to balance the shortcomings of one system with the advantages of the other system [23]. In this case study, ANFIS is utilized to build an efficient MPPT controller.

To design an MPPT controller using ANFIS, the first task is to gather the input–output data set for training purpose. This training data is generated using the PV model developed in previous case study. A step by step process of data generation is illustrated in the flowchart shown in Fig. 10. As a first step, values of the five unknown parameters for the considered PV panel and PV array are estimated using an efficient algorithm, e.g. similar to the one explained in the first case study in Sect. 3.1.

The training parameters are: N_{MAX} representing the number of training data points, (T_{MIN}, T_{MAX}) representing the minimum and maximum temperatures, and (S_{MIN}, S_{MA}) representing the minimum and maximum irradiation. T_{MIN}, T_{MAX}, S_{MIN}, and S_{MAX} can be specified depending upon the geographical location where the PV array is installed. The following step is to generate a random operating condition within the specified range and then modify the five parameters at this operating condition using Eqs. (18–23). The advantage of using the random operating condition is that it includes the uncertainties of the weather conditions within the training process. Then the transcendental nonlinear equation, Eq. (16), is solved for current calculation using an efficient numerical technique (e.g. Newton–Raphson

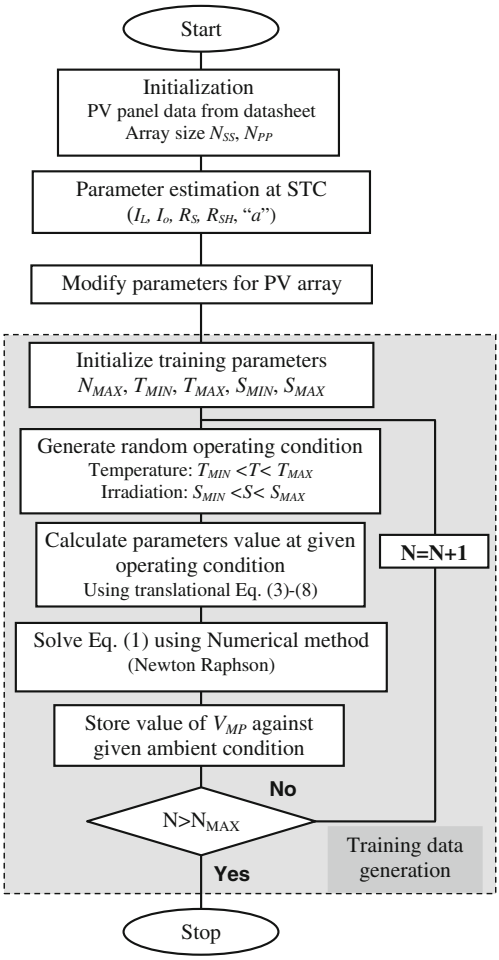


Fig. 10 Generation of input output data set for ANFIS training

in this case) and the value of the voltage corresponding to the maximum power point is stored against the specified operating condition. This process is executed for N_{MAX} times to generate the training data set of length N_{MAX}. After getting the input–output data set, the next step is to design the ANFIS-based MPPT by a hybrid learning algorithm. The parameters of the membership functions are adapted such that they track the input–output data finely.

The arrangement of the ANFIS-based MPPT controller is shown in Fig. 11. The input of the ANFIS-based MPPT controller is the ambient conditions, i.e. irradiation and temperature, and its output is the reference voltage (V_{REF}) which is normalized using the DC link voltage (V_{DC}). The normalized reference voltage ($V_{REF,\ norm}$) is fed back to the voltage control loop where a PI controller is used to maintain the output voltage of the PV array (V_{PV}) to the reference optimal voltage

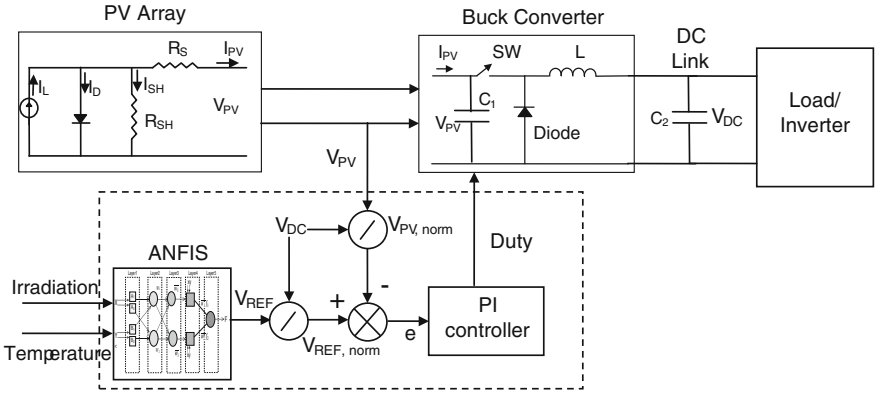


Fig. 11 PV system

Table 4 Specification PV panel parameter at STC

Panel parameter	Datasheet value	Model parameter	Estimated value
V_{OC} (V)	21.7	I_L (A)	3.35
I_{SC} (A)	3.35	I_0 (A)	1.7053e − 05
V_{MP} (V)	17.4	R_S (Ω)	0.00477
I_{MP} (A)	3.05	R_{SH} (Ω)	3.9601e + 04
N	36	a	1.78044

by adjusting the duty ratio of DC-DC converter, which results in a maximum power extraction.

3.2.2 Results and Discussion

A complete PV system used is shown in Fig. 11 and it consists of PV array, DC-DC buck converter, MPPT controller, DC link capacitor and inverter. The PV array will generate a varying DC voltage (V_{PV}) and current (I_{PV}) depending upon the weather conditions. The buck converter and DC link capacitor are utilized to smooth these varying quantities and make them accessible for the inverter. The MPPT is employed as a controller for the buck converter that adjusts the duty ratio of switch to extract maximum possible power from the PV array under all operating conditions.

The specifications of the PV system used are shown in Table 4. An array of 50 × 20 panels is used to show the operation of the ANFIS-based MPPT controller with a large PV power station. Buck converter is designed to work in a continuous conduction mode (CCM) and have the specifications of: $C_1 = 100$ μF, $L = 5$ mH, switching frequency of 5 kHz and DC link capacitor $C_2 = 500$ μF. The training parameters used to generate the set of input–output data set are: $N_{MAX} = 1000$,

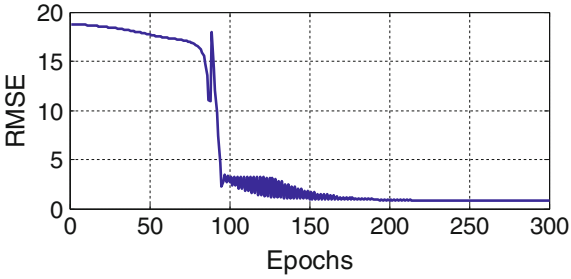


Fig. 12 Training error versus epochs for the ANFIS

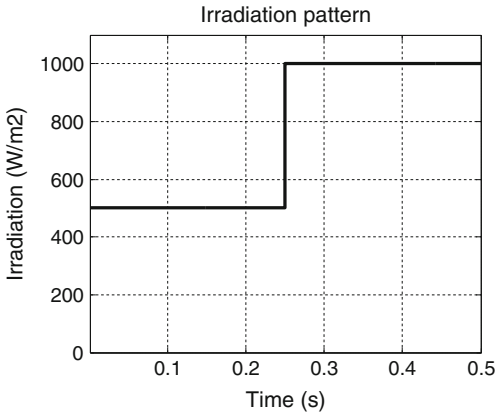


Fig. 13 Irradiation pattern

$T_{MAX} = 80$ °C, $T_{MIN} = 0$ °C, $S_{MAX} = 2000$ W/m^2, $S_{MIN} = 0$ W/m^2. These parameters show wide and dynamic range for temperature and irradiation that allows the designed MPPT to work efficiently under uncertain operating conditions.

The ANFIS-based MPPT is developed in MATLAB/Simulink using three generalized bell (g-bell) shaped membership functions. These membership functions are selected by comparing the training root mean square error (RMSE) for different number of membership functions. A hybrid learning algorithm is utilized that uses the LSE to adapt the consequent parameters and a back-propagation method to optimize the premise parameters of the membership functions. The epochs selected for training purpose is 300 that reduced the RMSE to a lower value of 0.8 as shown in Fig. 12.

To show the effectiveness of the ANFIS-based MPPT, a comparison with the conventional incremental conductance (InCond) method is demonstrated. Both simulation and experimental tests are conducted under the step-up change in irradiation level and is shown in Fig. 13. It shows that the irradiation is constant with a value of 500 W/m^2 up to 0.25 s and then increases drastically to 1,000 W/m^2. The P–V curves for the selected PV array under low (500 W/m^2 and 25 °C) and normal (1,000 W/m^2 and 25 °C) irradiation levels are shown in Fig. 14. It can be seen that the maximum power that can be generated by the PV array at low irradiation level is

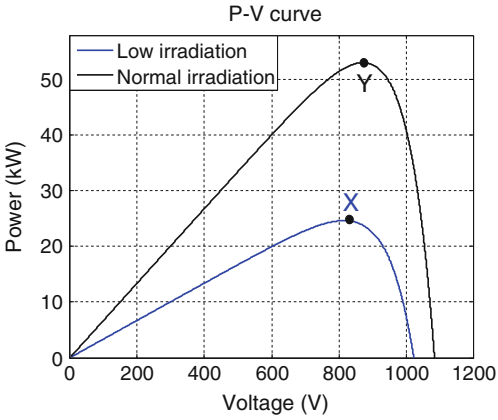


Fig. 14 PV curve under normal and low irradiation conditions

24.669 kW and is labeled as point X on the graph. After a step-up change in irradiation, the operating point shifts to point Y (point at STC) having the maximum possible power of 53.07 kW ($V_{MP} * I_{MP} * N_{SS} * N_{PP} = 53.07$ kW).

A. Simulation Results

A nonlinear time domain simulation is carried out with the ANFIS-based and InCond-based MPPT controllers. For InCond method, a fixed value of perturbation step size is chosen based on a tradeoff between fluctuations in the steady state and tracking speed. Figure 15 shows a comparison of the PV array power output (P_{PV}) for both the controllers. It can be seen from the graph that the MPP reached by the ANFIS-based MPPT controller is much faster than the conventional InCond in both the tracking regions, i.e. start of the algorithm and step-up change in irradiation. For example, it can track the MPP in 0.04 s while the InCond takes 0.1 s in the start of the algorithm. When a step-up change in irradiation occurs, the MPPT controller shifts the operating point from point X to Y and it can be noticed from the plot that the response of InCond is much slower than the proposed ANFIS-based MPPT.

To demonstrate the efficiency of the proposed controller in the steady state region, a portion of the graph from 0.3 to 0.34 s is enlarged where irradiation level is maintained at 1,000 W/m^2. It can be seen from the enlarged graph that the ANFIS-based MPPT has smoother response as compared to InCond which shows a considerable amount of fluctuation in the steady state and its power varies from 53.07 to 52.78 kW that shows oscillation of 294 W. Although it is possible to diminish these fluctuations by reducing the perturbation step size, it results in even much slower tracking of MPP. The simulation results for the duty cycle PV array voltage (V_{PV}) and current (I_{PV}) are shown in Figs. 16 and 17. This verifies the effectiveness of the proposed MPPT under the changing irradiation condition.

From the simulation results, it can be inferred that the proposed ANFIS-based controller is faster than InCond controller in transitional state, and it has fewer

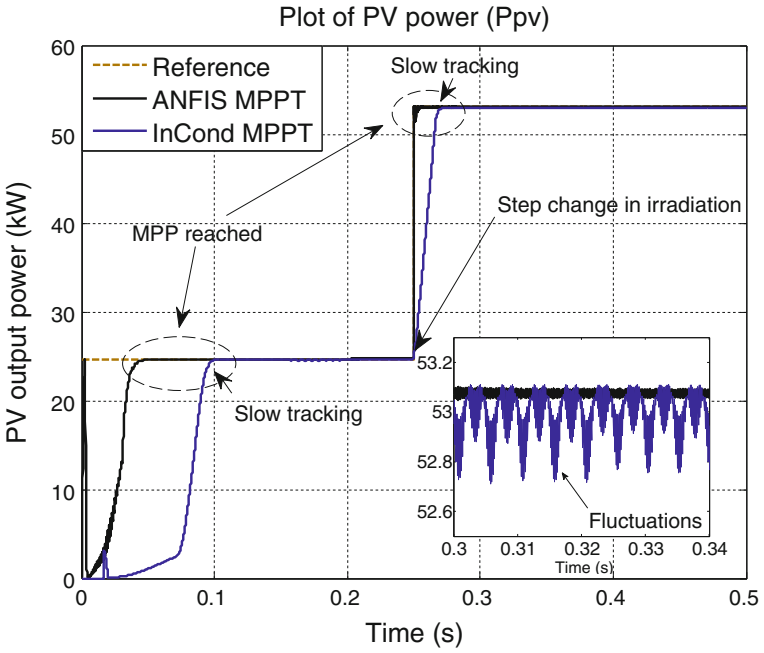


Fig. 15 Characteristics of PV output under step-up irradiation change

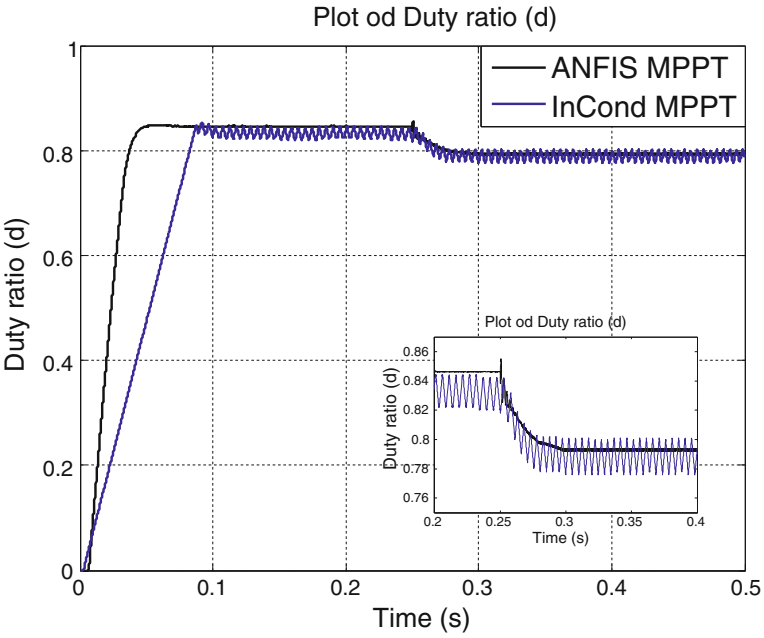


Fig. 16 Duty ratio under step-up irradiation change

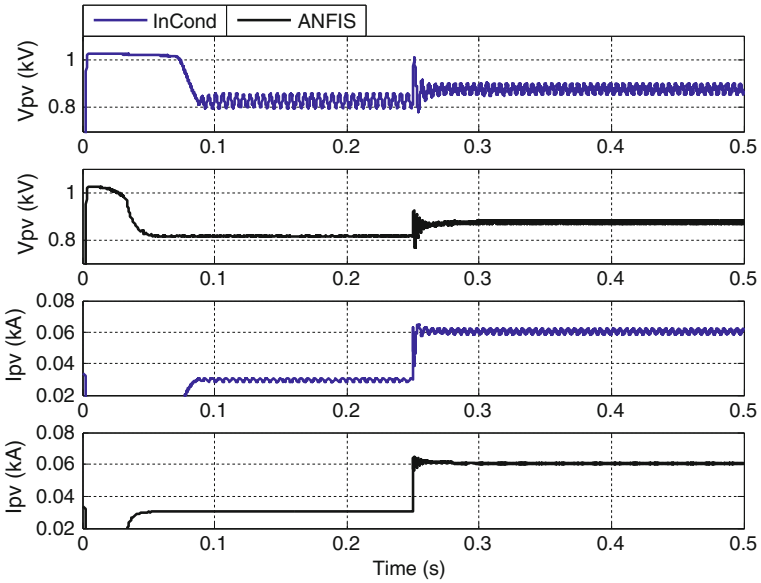
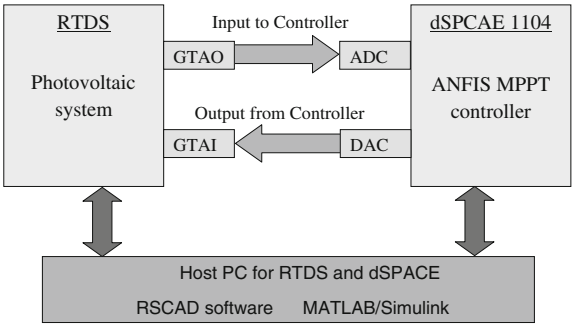


Fig. 17 Characteristics of PV voltage (VPV) and current (IPV)

Fig. 18 Closed-loop control system



oscillations in the steady state. All these factors will cause less power loss and result in more power output from the PV array.

B. Experimental Results

The effectiveness of the proposed ANFIS-based MPPT controller is experimentally verified using the Real-Time Digital Simulator (RTDS) and dSPACE controller. A complete PV system is developed in RTDS and the ANFIS-based MPPT is designed in dSPCAE controller and integrated with RTDS as shown in Fig. 18.

The same disturbance as used in the simulation test (step-up change in irradiation level) is applied. The experimental results for PV power (P_{PV}), PV voltage (V_{PV}), PV current (I_{PV}) duty cycle are shown in Figs. 19 and 20 for both controllers.

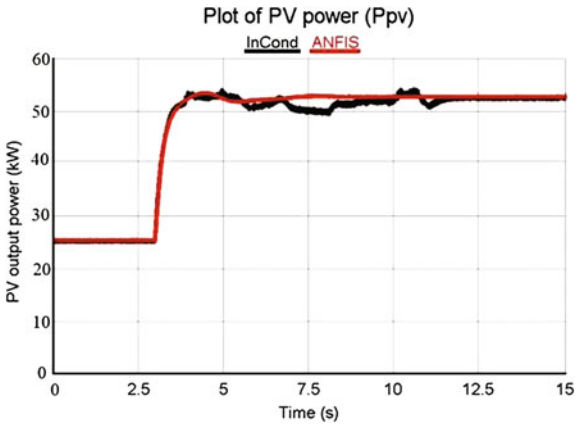


Fig. 19 Experimental waveform for PV power output

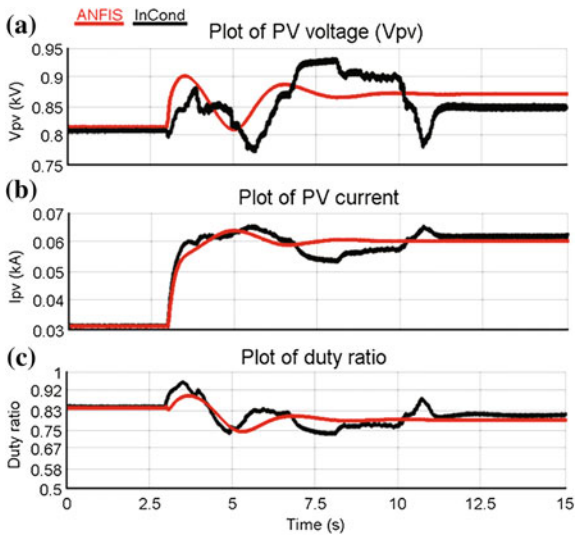


Fig. 20 Experimental waveform for PV voltage, PV current and duty ratio

It can be seen that the ANFIS-based MPPT controller has less fluctuations and it can reach the steady state faster than the conventional InCond method. This experimentally verifies the competence of the ANFIS-based MPPT over the conventional method.

C. Comparison of Simulation and Experimental Results

The results from the MATLAB/Simulink simulations are compared with the experimental results to explore the validity of the ANFIS-based MPPT controller. A comparison of the system response and performance for the disturbance under discussion is shown in Figs. 21, 22 and 23. Figure 21 depicts the PV power output (P_{PV}) and shows how the proposed controller track the MPP in MATLAB/Simulink and experimental simulations under the step-up change in irradiation level.

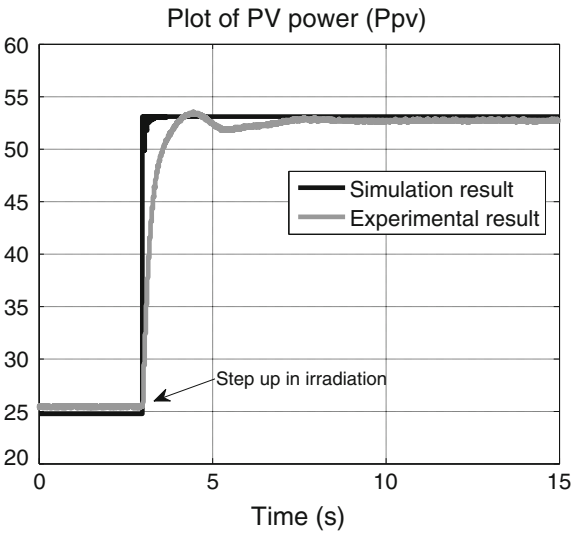


Fig. 21 Comparison of PV power (P_{PV})

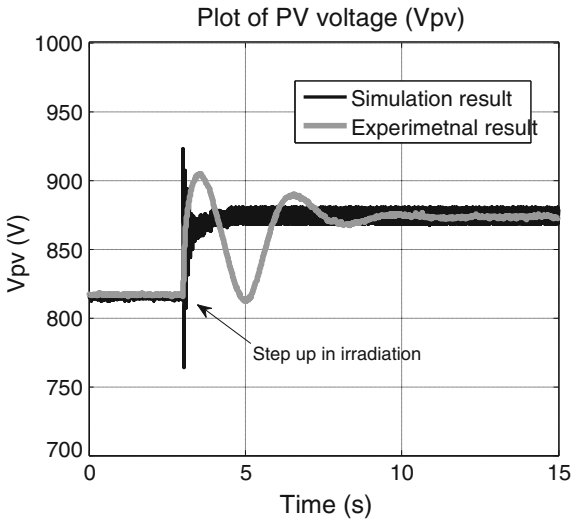


Fig. 22 Comparison of PV voltage (V_{PV})

Comparisons of the PV output voltage (V_{PV}) and current (I_{PV}) are illustrated in Figs. 22 and 23, respectively. It can be seen that the experimental results are very much similar to the simulation results. Both the experimental and MATLAB/Simulink results validate the accuracy of the ANFIS-based MPPT controller model.

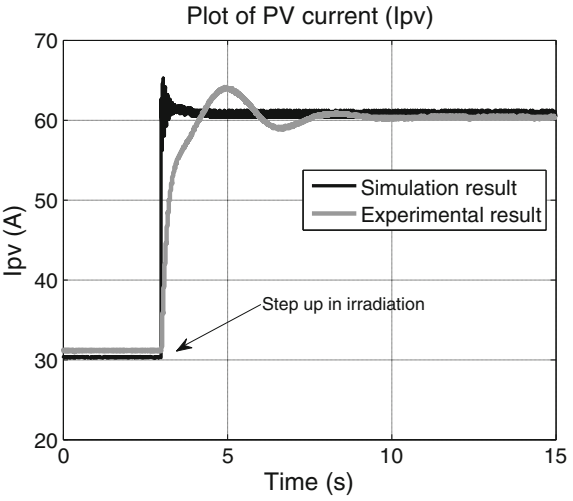


Fig. 23 Comparison of PV current (I_{PV})

4 Directions for Future Work

- The following subjects are recommended for future work.
- The developed PV simulator can be modified by using two-diode PV model that will add two additional unknown parameters to optimize and comparison can be done between modified and proposed simulator to investigate their efficiencies.
- Different optimization techniques can be used instead of DE to estimate the model parameters of PV, to study which optimizing technique is better.
- Different objective function can be used to test the efficiency of the evolutionary techniques.
- Partial shading condition is one of the major issues and causes multiple peaks in the PV curve and made it difficult to track the global MPP. The proposed ANFIS-based MPPT controller is designed for uniform irradiation condition and it can be improved to work in the partial shading conditions.
- Effectiveness of the ANFIS-based MPPT can be tested under actual real data of irradiation and temperature.
- The developed PV system model and ANFIS-based MPPT controller can be interfaced with the power grid through inverter and effects of changing environmental conditions on power grid can be studied.
- Practical setup of actual PV array with all necessary sensors and controllers can be installed to test the performance of proposed ANFIS-based MPPT controller.

Acknowledgments The authors would like to acknowledge the support provided by King Abdulaziz City for Science and Technology (KACST) through the Science & Technology Unit at King Fahd University of Petroleum & Minerals (KFUPM) for funding this work through projects No. 11-ENE1632-04 and 11-INF1658-04 as part of the National Science, Technology and Innovation Plan.

References

1. European Technology Platform SmartGrids, Strategic Deployment Document for Europe's Electricity Networks of the Future, Third Report, http://www.smartgrids.eu/documents/SmartGrids_SDD_FINAL_APRIL2010.pdf. Accessed April 2010
2. F.B. Beidou, W.G. Morsi, C.P. Diduch, L. Chang, Smart grid: challenges, research directions and possible solutions, in *Proceedings of the 2nd IEEE International Symposium on Power Electronics for Distributed Generation Systems* (2010), p. 670
3. F. Li, W. Qiao, H. Sun, H. Wan, J. Wang, Y. Xia, Z. Xu, P. Zhang, Smart transmission grid: vision and framework. IEEE Trans. Smart Grid **1**(2), 168–177 (2010)
4. Y. Wang, D. Ruan, J. Xu, M. Wen, L. Deng, in *Computational Intelligence Algorithms Analysis for Smart Grid Cyber Security*, eds. by Y. Tan, Y. Shi, K.C. Tan. ICSI 2010, Part II, (LNCS 6146, 2010), p. 77
5. G.K. Venayagamoorthy, Dynamic, stochastic, computational, and scalable technologies for smart grids. IEEE Comput. Intell. Mag. **6**(3), 22–35 (2011)
6. G. Ye, R. Rao, M. Li, A multiobjective resources scheduling approach based on genetic algorithms in grid environment, in *Proceedings of the 5th International Conference on Grid and Cooperative Computing Workshops (GCCW '06)*, (2006)
7. F. Xhafa, A. Abraham, Computational models and heuristic methods for grid scheduling problems. Future Gener. Comput. Syst. **26**(4), 608–621 (2010)
8. R.G. Harley, J. Liang Computational intelligence in smart grids, in *Proceedings of the IEEE Symposium Series on Computational Intelligence*, (2011)
9. F.D. Angelis, M. Boaro, D. Fuselli, S. Squartini, F. Piazza, in *A Comparison Between Different Optimization Techniques for Energy Scheduling in Smart Home Environment*. Neural Nets and Surroundings, Smart Innovation, vol 19 (Systems and Tech, 2013), pp. 311–320
10. Y. Yare, G.K. Venayagamoorthy, U.O. Aliyu, Optimal generator maintenance scheduling using a modified discrete PSO. IET Proc. Gener. Trans. Distrib. **2**(6), 834–846 (2008)
11. M.A. Abido, Multiobjective evolutionary algorithms for electric power dispatch problem. IEEE Trans. Evol. Comput. **10**(3), 315–329 (2006)
12. M. Hassan, M.A. Abido, Optimal design of microgrids in autonomous and grid-connected modes using particle swarm optimization. IEEE Trans. Power Electron. **26**(3), 755–769 (2011)
13. M.H. Khan, M.A. Abido Optimizing the controller of PV system to enhance the dynamic stability of smart grid network, in *Proceedings of the Third International Conference on Intelligent Systems Modeling and Simulation (ISMS 2012)*, (Kota Kinabalu, 2012), p. 96
14. B. Luitel, G.K. Venayagamoorthy, Wide area monitoring in power systems using cellular neural networks. in *Proceedings of the IEEE Symposium on Computational Intelligence Applications in Smart Grid (CIASG)*, (2011)
15. Z. Vale, G.K. Venayagamoorthy, J. Ferreira, H. Morais, in *Computational intelligence applications for future power systems*, eds. by A. Madureira et al. Computational Intelligence for Engineering Systems: Emergent Applications, Intelligent Systems, Control and Automation: Science and Engineering, (Springer, New York, 2011)
16. M. Sheraz, M.A. Abido An efficient MPPT controller using differential evolution and neural network, in *Proceedings of the IEEE International Conference on Power and Energy (PECon)*, (2012), p. 378
17. P.J. Werbos, Computational intelligence for the smart grid-history, challenges, and opportunities. IEEE Comput. Intell. Mag. **6**(3), 14–21 (2011)
18. Z. Jiang, *Computational Intelligence Techniques for a Smart Electric Grid of the Future*, eds. by W. Yu, H. He,N. Zhang. ISNN 2009, Part I, (LNCS 5551, 2009), p. 1191
19. G.K. Venayagamoorthy, Potentials and promises of computational intelligence for smart grids. *IEEE Power and Energy Society General Meeting* (2009)

20. M. Carvalho, C. Perez, A. Granados, An adaptive multi-agent-based approach to smart grids control and optimization. Energy Syst. **3**, 61–76 (2012)
21. M. Peters, W. Ketter, M. Saar-Tsechansky, J. Collins, A reinforcement learning approach to autonomous decision-making in smart electricity markets. Mach. Learn. **92**, 5–39 (2013)
22. R. Storn, K. Price, Differential evolution—A simple and efficient heuristic for global optimization over continuous spaces. J. Global Optim. **11**(4), 341–359 (1997)
23. J.S.R. Jang, ANFIS: adaptive-network-based fuzzy inference system. IEEE Trans. Syst. Man Cybern. **23**(3), 665–685 (1993)
24. J.S.R. Jang, Neuro-fuzzy modeling and control. Proc. IEEE **83**(3), 378–406 (1995)
25. W. Kim, W. Choi, A novel parameter extraction method for the one-diode solar cell model. Sol. Energy **84**(6), 1008–1019 (2010)
26. M.G. Villalva, J.R. Gazoli, E.R. Filho, Comprehensive approach to modeling and simulation of photovoltaic arrays. IEEE Trans. Power Electron. **24**(5), 1198–1208 (2009)
27. N. Moldovan, R. Picos, E. Garcia-Moreno, Parameter extraction of a solar cell compact model using genetic algorithms. in *Proceedings of the Spanish Conference on Electron Devices* (2009), p. 379
28. M. Ye, X. Wang, Y. Xu Parameter extraction of solar cells using particle swarm optimization. J. Appl. Phys. **105**(9), 094502–094502-8 (2009)
29. K. Ishaque, Z. Salam, A review of maximum power point tracking techniques of PV system for uniform insolation and partial shading condition. Renew. Sustain. Energy Rev. **19**, 475–488 (2013)
30. K. Kobayashi, H. Matsuo, Y. Sekine, A novel optimum operating point tracker of the solar cell power supply system. in *Proceedings of the 35th IEEE Annual Power Electronics Specialists Conference* (2004), p. 2147
31. M. Fortunato, A. Giustiniani, G. Petrone, G. Spagnuolo, M. Vitelli, Maximum power point tracking in a one-cycle-controlled single-stage photovoltaic inverter. IEEE Trans. Industr. Electron. **55**(7), 2684–2693 (2008)
32. B.N. Alajmi, K.H. Ahmed, S.J. Finney, B.W. Williams, Fuzzy-logic-control approach of a modified hill-climbing method for maximum power point in microgrid standalone photovoltaic system. IEEE Trans. Power Electron. **26**(4), 1022–1030 (2011)
33. A.K. Abdelsalam, A.M. Massoud, S. Ahmed, P.N. Enjeti, High-performance adaptive perturb and observe MPPT technique for photovoltaic-based microgrids. IEEE Trans. Power Electron. **26**(4), 1010–1021 (2011)
34. F. Liu, S. Duan, F. Liu, B. Liu, Y. Kang, A variable step size INC MPPT method for PV systems. IEEE Trans. Industr. Electron. **55**(7), 2622–2628 (2008)
35. K. Ishaque, Z. Salam, M. Amjad, S. Mekhilef, An improved particle swarm optimization (PSO)–based MPPT for PV with reduced steady-state oscillation. IEEE Trans. Power Electron. **27**(8), 3627–3638 (2012)
36. M. Sheraz, M.A. Abido An efficient fuzzy logic based maximum power point tracking controller for photovoltaic systems. in *Proceedings of the International Conference on Renewable Energies and Power Quality (ICREPQ'13),* (Bilbao, 2013)

On-line Demand Management of Low Voltage Residential Distribution Networks in Smart Grids

Farhad Shahnia, Michael T. Wishart and Arindam Ghosh

Abstract A novel intelligent online demand management system is discussed in this chapter for peak load management in low voltage residential distribution networks based on the smart grid concept. The discussed system also regulates the network voltage, balances the power in three phases and coordinates the energy storage within the network. This method uses low cost controllers, with two-way communication interfaces, installed in costumers' premises and at distribution transformers to manage the peak load while maximizing customer satisfaction. A multi-objective decision making process is proposed to select the load(s) to be delayed or controlled. The efficacy of the proposed control system is verified by a MATLAB-based simulation which includes detailed modeling of residential loads and the network.

Keywords Smart grid · Demand management · Peak load shaving · Voltage control · Power balancing · Decision making

F. Shahnia (✉)
Department of Electrical and Computer Engineering, Center of Smart Grid and Sustainable Power Systems, Curtin University, Perth, Australia
e-mail: farhad.shahnia@curtin.edu.au

M. T. Wishart
Technology Development Department, Ergon Energy, Brisbane, Australia
e-mail: mike.wishart@ergon.com.au

A. Ghosh
School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane, Australia
e-mail: a.ghosh@qut.edu.au

Z. H. Khan et al. (eds.), *Computational Intelligence for Decision Support in Cyber-Physical Systems*, Studies in Computational Intelligence 540,
DOI: 10.1007/978-981-4585-36-1_10,

1 Introduction

Distribution networks are designed to supply peak loads to ensure acceptable reliability, despite the fact that these peak loads typically occur for a small fraction of the year [1]. Therefore, the overall electricity infrastructure cost is largely determined by the peak load on the network; hence, there is a strong motivation to minimize peak load growth. In many parts of the world, peak load growth in residential areas is higher than the consumption growth. As an example, in Queensland Australia, electrical utilities Energex and Ergon traditionally experienced an average annual residential peak load growth of 7–13 % compared with an annual residential consumption growth of 2–3 % due to a number of factors including the proliferation of air-conditioning [2, 3]. This has resulted in large annual capital expenditures on system upgrades. In the future, the introduction of plug-in electric vehicles (PEVs) is expected to further increase the peak load especially in residential areas [4–6].

Much work has been historically done on direct or indirect demand management [7–12]. Direct demand management schemes, often called Direct Load Control (DLC) systems, typically make use of a control signal from the utility to directly control loads. The water heater ripple control systems currently used in many parts of the world are an example of a traditional DLC system [13, 14]. Recent schemes often propose using real-time pricing as the control signal to trigger automated action from home automation controllers [15, 16]. Some pilot projects are also being conducted of individual house load management [17]. In addition, direct control methodologies for PEV charging have recently been proposed to minimize the impact of PEV charging [18, 19]. Indirect demand management schemes use price as a control variable to influence consumers' behavior and thus indirectly control the load. For example, time of use tariffs typically increase the price of power during peak periods thus encouraging consumers to shift their consumption to off-peak period [20–22].

On the other hand, keeping the voltage in distribution networks within the standard acceptable range is of high importance and often a statutory requirement for utilities. The voltage variation in low voltage distribution networks depends on different factors including load variation and quantity of rooftop photovoltaic cells (PVs) within residential premises. Utilities traditionally design the low voltage circuitry to limit voltage drop under peak load conditions. This may result in a methodology of setting distribution transformer tap settings to relatively high values to allow for the voltage drop along the length of the low voltage distribution feeder in combination with using relatively low impedance conductor with high cross-sections [1, 23]. However, this methodology may result in a voltage rise problem during off-peak periods in networks with high penetration level of PVs. For preventing this issue, some utilities have placed a limit on the penetration level of PVs in low voltage distribution networks. For example, Energex does not permit a new household to install rooftop PV if penetration level in the feeder is above 30 %.

The grid-connected rooftop PVs were financially beneficial for the householders because of the high electricity feed-in tariff. However, it is expected that the feed-in tariffs will be stopped soon by many governments. This will provide some motivation for householders to install batteries to store the generated power from PVs for later consumption instead of injecting it into grid for low financial reward. This is favorable for reducing the voltage rise problems in network during off-peak periods and can lead to a higher level of PV penetration in the network in the future [24]. In a similar way, they can be helpful in the network voltage drop management.

On the other hand, phase balancing is also of high interest in distribution networks. Although the utilities aim to minimize the unbalance among the three phases by equally distributing houses among the three phases, in practice many houses have historically been connected in an ad-hoc manner resulting in unbalanced connection. In addition, different power consumption patterns in houses, house sizes and network configurations have high impact on unbalance and this is further affected by PV, PEVs and batteries penetration. Different methods are already proposed for unbalance improvement in networks by installation of power electronics-based devices [25]. However, unbalance can also be targeted when controlling the loads in the low voltage residential distribution networks using the smart demand management.

In this chapter, a smart on-line direct demand management system is discussed for low voltage distribution networks. The main objective of the discussed novel method is to prevent overloading the distribution and upstream transformers at peak load periods. This method works based on instantaneous load levels and requires installation of smart controllers with local communications capability. Further, this smart demand management system is modified to consider voltage drop and unbalance when controlling the loads.

The load control decision during peak periods (i.e. which load to delay or control) is a complex process, depending on multiple different criteria such as customer priority, load operation satisfaction, overloading level and number of loads to be controlled. In this chapter a Multi-Objective Decision Making (MODM) method [26] is utilized to achieve the load control objective with minimal impact on customers. MODM method is developed with the following objectives:

- Maintain the transformer loading below its nominal rating,
- Balance the three phases,
- Maintain network voltage within standard limits,
- Coordinate battery charging and discharging in the network.

Section 2 presents the low voltage residential distribution network, under consideration in this chapter. In Sect. 3, the proposed control method is described along with its required hardware, control algorithms. Section 4 discusses the decision making process for load control. In Sect. 5, the analysis methodology is described including detailed modeling of different residential loads and load flow

analysis. The simulation results for the network under consideration, with and without the novel smart demand management control system, are provided in Sect. 6. The effectiveness of the system in controlling the peak demand, network voltage drop and power unbalance among the three phases, while maintaining customer satisfaction, is demonstrated through simulation results.

2 Network Under Consideration

Figure 1 shows the schematic of a typical low voltage residential distribution feeder in a suburban area. It is assumed that a 100 kVA, 11 kV/400 V distribution transformer supplies residential loads in the feeder. The schematic diagram of the novel smart demand management system is also shown in this figure.

The network assets are traditionally designed for a certain peak load level based on standardized After Diversity Maximum Demand (ADMD) figures [1]. For example, Ergon Energy, uses an ADMD of 4–5 kVA per detached dwelling for suburban houses, based on historical peak load data and obviously neglects PEVs. Transformer and conductor ratings and transformer tap settings are calculated using the ADMD values along with the associated voltage drop and consideration of the voltage drop on the Medium Voltage (11 kV) network.

In practice, the peak load value in residential networks is season and area dependent but generally occurs in the evening. Uncontrolled Level I charging of PEVs (from a house outlet at 230 V 10/15 A) is expected to be added to this peak since it is expected that customers will generally plug in their vehicles for charging as they return home. If the traditional network design methodology is followed, penetration of PEVs will potentially have significant implications on the network infrastructure cost.

3 Smart Demand Management System

The main requirements for such a smart demand control system are:

- Effectiveness: the system should limit peak load while assuring customer satisfaction.
- Low cost: the system should utilize low cost hardware.
- Scalable: the system should be easily scalable to larger networks.
- Robustness: the system should be fault tolerant.

Considering the distribution network depicted in Fig. 1, the main objective of the control system is to ensure that the transformer and conductors do not exceed their ratings while minimizing the negative impact on consumers. In addition, the

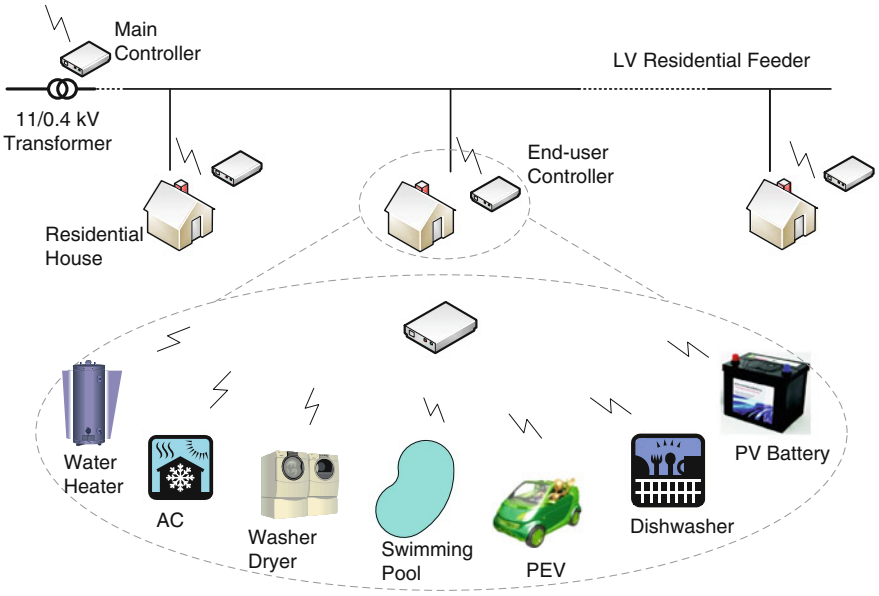


Fig. 1 Schematic diagram of a low voltage residential distribution feeder with the discussed novel smart demand management system

control system might be utilized to improve voltage drop as well as power unbalance among the three phases.

It is proposed that a micro-processor based intelligent controller, denoted end-user controller, is installed in each house to measure and control loads and to communicate with the controller located at the transformer. All controllers have low-bandwidth two-way communication capability. The end-user controllers measure the power consumption of each controllable device in the house and send that information to the relevant transformer controller every 2 min. The schematic diagram of the novel smart demand management system is shown in Fig. 1.

When the transformer controller detects an overload in the transformer, it decides which load(s) should be controlled and sends a command to the end–user controller in the related house to delay/adjust that load(s). In this analysis, eight controllable devices are considered within residential premises. These controllable devices are swimming pool pumps, electric water heaters, electric vehicle chargers, hysteresis-type Air Conditioners (ACs), inverter-type ACs, clothes washers, dish washers and clothes dryers. In addition, charging/discharging of residential batteries can be controlled. The load selection procedure is discussed in the rest of this section. After the selected load is delayed or adjusted, the end-user controller sends a confirmation command to the transformer controller. Upon receiving this, the transformer controller re-measures the total transformer loading. If the transformer loading is now below the threshold, no further action is taken. If the loading still exceeds the threshold, the transformer controller will again choose

another load to be delayed/adjusted. This process continues until all controllable loads are delayed/adjusted.

The specific control action depends on the load type. Most loads (e.g. clothes/dish washers, clothes dryers, swimming pool pumps, electric water heaters and electric vehicle chargers) are delayed by 15 min and then reconnected, while the set point of locally controlled loads such as inverter ACs and water heaters are adjusted for 15 min and then reset.

Various incentive methods/tariffs can be considered to encourage the householder to participate in this control scheme. While not the focus of this chapter, this could be as simple as free installation of controllers along with rebates or discounts on electricity bills.

In the discussed system, the control signals are generated as soon as the load reaches the steady state transformer rating. However, this rating limit can also be made dynamic based on a transformer thermal model which considers the transformer rating, the overload magnitude and period, and the ambient temperature [27]. This would generally allow a greater load before limiting, while still maintaining the transformer overload protection.

Reducing the total loading of a network will indirectly result in improved voltage profile within the network. However, there is no certainty that if the load is below its limit, the low voltage network voltage profile falls within its limits. For improving the effectiveness of the controller, other objectives are considered in addition to transformer loading control. So far, it was assumed that the controller starts action if the transformer loading is above its nominal rating. Now, let us assume that the controller starts action if the transformer loading is above its nominal rating or the voltage at any node along the feeder is below the standard voltage limit. For this purpose, it is necessary that the end-user controllers measure the input voltage rms of each house in addition to the power consumption of each device in the house and send that information to the transformer controller every 2 min. Therefore, when the transformer controller detects an overload in the transformer or non-standard voltage drop along the feeder, it decides which load(s) should be controlled and sends a command to the end-user controller in the related house to delay/adjust that load. After the selected load is delayed or adjusted, if the transformer loading still exceeds its threshold or voltage drop exceeds standard limits at any bus along the feeder, the transformer controller will again choose another load to be delayed or adjusted. This process continues until all controllable loads are delayed or reduced.

3.1 Voltage Control Algorithm

Different methods can be used for regulating the voltage in network peak period. In the first method, the controller can be scheduled to adjust/delay a load if the voltage along the feeder drops below the standard voltage limit at any node. However, the voltage improvement on the network is dependent on the location of

the controlled load and the required voltage drop will not necessarily be achieved even after controlling many loads.

More effective results can be achieved if transformer maximum rating level is reduced in case of excessive voltage drop along the feeder. Later in this chapter, for the hypothetical feeder considered, it is shown that the voltage drop along a feeder can be kept above 0.94 pu along the feeder if the transformer maximum overloading limit is reduced to 80 % of its nominal rating. The 80 % limit was calculated after several observations and choosing different levels between 75–95 % for the network under consideration. It is to be reminded that the best limit can be chosen based on the number of loads controlled and the time period that the non-standard voltage exists in the network.

3.2 Control of Batteries

Different charging and discharging control algorithms can be considered for batteries connected to PVs in residential premises. Generally householder owned batteries would be used to optimize householder value, for example to minimize the exported power to the grid when feed-in tariffs are not attractive. In this study however, it is assumed that the batteries are controlled for utility purposes. It is assumed that they automatically start to charge from their arbitrary initial charge to 100 %, with a constant or variable rate, when the PV output power is available. Hence, the PV and battery combination can act such that all active power generated from PVs gets charged into batteries until they are fully charged. Once the batteries are fully charged any PV generated power is fed into the grid.

There are two general strategies to control the battery discharge. In the first strategy, the batteries will start to discharge to prevent transformer overloading. In this case, batteries will start to discharge earlier in the day and the network will suffer from voltage drop problem later in the day after the batteries are discharged to their minimum acceptable SOC. Hence, the transformer rating limitation has to be applied for a longer time per day, which is not desired.

In the second strategy, the batteries will start to discharge to prevent non-standard voltage drop in network. Therefore, in this study, discharging of batteries is used for voltage support and regulation in the network. Consequently, they do not start to discharge if the voltages are above standard voltage drop limit, irrespective of transformer loading.

Therefore, in the discussed system, a voltage regulation strategy is considered as the main criterion for controlling the discharge of the batteries. In this strategy, two options can be considered as discussed below:

- The first option is the stand-alone control of each battery. In this way, each battery will start to discharge with a constant rate if the input voltage of the house, in which the battery is installed, goes below the standard voltage limit. Although this decentralized control method is easy to develop and is

communication free, it might not be useful for the batteries installed at houses at the beginning of the feeder as the voltage in those nodes might never go below standard voltage drop limit even during network peak periods.
- The better option is that all batteries within the network start to discharge when there is a non-standard voltage drop problem at any bus in any of the phases within the network. This simultaneously addresses both the voltage drop and the transformer overloading issues. Hence less control commands are used for the loads resulting in better customer satisfaction. This is the selected option in this study for discharging control of rooftop PV batteries in the network.

4 Multi-objective Decision Making Process

MODM is a methodology for selecting the best action (or decision) from a large number of alternatives given multiple (often competing) objectives. A decision matrix is used to weight different alternatives and derive the resulting action. A survey of different MODM techniques is provided in [26].

In the discussed system, eight controllable residential load types are considered as alternatives for delay or control by the distribution transformer controller. In addition, there are different criteria for defining which load(s) to be controlled. Using the MODM method, it is possible to consider the effect of several criteria, each with a different weighting value on prioritizing the loads to be delayed or adjusted. The criteria which have effect on prioritizing the loads are:

1. *User Priority*
   Customers may set the general priority of load delay or adjustment i.e. the order in which their appliances are delayed or controlled during peak load periods. The user priority is converted to a numerical value in range of [0 1] for their eight controllable devices, where 1 and 0 show the highest and lowest priority for delay/control, respectively.

2. *Flexibility of Device Operation*
   Inherent characteristics of different load types result in different flexibilities for disconnection and reconnection of appliances. For example, a swimming pool pump can work any time of the day if it satisfies its desired total hours of operation per day. Therefore, it has a high flexibility. Clothes/dish washers have a lower flexibility. This is due to this fact that if they are disconnected during a cycle, the heated water inside will cool down; needing to be warmed up again next time it starts. Water heaters and ACs are considered to have a high flexibility (subject to their satisfaction criterion discussed next).

3. *Satisfaction Index*
   A satisfaction index is defined to represent how close a device/appliance is to its optimum state of operation. This index is dynamic and is updated every 5 min. This index is calculated differently for each load as shown schematically in Fig. 2. They are briefly discussed below:

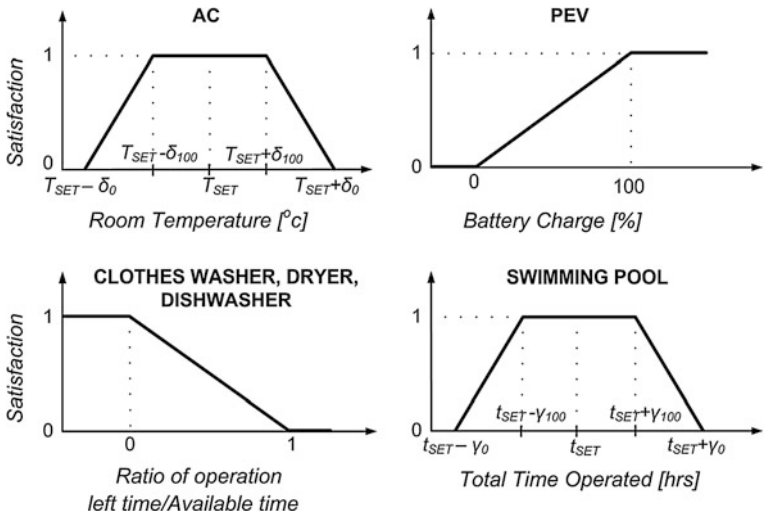


Fig. 2 Satisfaction index calculation for different type of residential loads

- AC satisfaction index depends on how close the room temperature is to its set point.
- PEV satisfaction depends linearly on the state of charge of its battery.
- Clothes washer, dryer and dishwasher satisfaction depends on the ratio of remaining operational time and available time. Available time is based on constraints set by the user i.e. washing/drying must be finished by a certain time.
- Swimming pool satisfaction depends on total operational time in the last 24 h compared with set time.

4. *Power Matching*
   It is more desirable for the transformer controller to select a load which closely matches the required power decrease than one which is highly dissimilar. Therefore, the controller calculates a dynamic power similarity criterion for each load at each decision making step. All loads are normalized in range of [0 1] where 1 shows the power consumption of a specific load is very close to the required power reduction.

5. *High Power Consumption*
   Assuming all customers are on the same tariff, it is desirable for the controller to first control an appliance from the house with higher total electric power consumption rather than with lower total power, since the houses with the highest consumptions are the biggest contributors to the overload. Therefore, at each decision making step, a numerical value in range [0 1] is allocated for all controllable devices in each house expressing the ratio of total power of the house compared to other houses, where 1 and 0 show the house respectively with highest and lowest total power consumption at that moment.

6. *Power Balancing*
   Power balancing is of high interest in distribution networks. It can be considered as a criterion in the discussed smart demand management system when controlling the loads within the network. For this purpose, at each decision making step, a numerical value in range [0 1] is allocated for all controllable devices in all the houses connected to one phase. This value will be high if the houses are connected to a higher loaded phase and will be low if the houses are connected to the lowly loaded phase. The difference between high and low values is normalized to demonstrate the power unbalance among the three phases.
   It is to be advised that, this should be only applied to the networks with equal number of houses in each phase where the unbalance is mostly because of the different power consumptions in houses not number of houses in each phase. However, there are some cases, that the utilities have not distributed the houses equally among three phases. In this case, one phase might be always more loaded than other phases due to higher number of houses connected to that. In this situation, applying the proposed control method is not fair as it will always result in more control commands to the houses connected to the highly loaded phase while less control commands will be sent to houses in other phases. This method can be applied for such systems taking into account the number of houses in each phase in the weightings.

7. *Participation of Batteries*
   Existence of batteries in a house and their participation in smart demand management system, as mentioned before, is helpful for reducing voltage problem and transformer overloading in the network. Therefore, for encouraging the householders with rooftop PVs to install batteries, an incentive method can be implemented such that lower load disconnection/adjustment priority will be applied for the houses with installed batteries in comparison to the houses without batteries. This numerical value is in range [0 1]. Also this criterion in not compulsory but utilities can adopt it for encouraging householders to install batteries and participate their battery control in the smart demand management system. Rewards such as tariff discounts or paybacks, similar to existing rewards for smart ACs, water heaters and swimming pool pumps by Energex, can be utilized for this purpose [28].

8. *Electricity Price*
   The MODM approach is ideally suited for the inclusion of time varying wholesale energy market costs of primary interest to energy retailers. Thus if purchased wholesale power cost is added to the decision making process, energy retailers can use this scheme to shift some of the network load to periods with lower energy cost. Hence, this will result in reducing the generation component of electricity bills, thus creating downward pressure on electricity prices and indirectly economically benefiting the customers. This customer economic benefit mechanism can be as simple as a lower tariff if they accept this load control scheme.

Table 1 Decision making matrix for load selection within smart demand management system

House no.	Criteria weightings		Priority B_1	Flexibility B_2	Satisfaction B_3	Power matching B_4	High power B_5	Power balance B_6	Battery B_7	Control priority value
	Alternatives									
1	1	Water heater	$H(1, 1)$	$H(1, 2)$	$H(1, 3)$	$H(1, 4)$	$H(1, 5)$	$H(1, 6)$	$H(1, 7)$	D_1
	2	Swimming pool	$H(2, 1)$	$H(2, 2)$	$H(2, 3)$	$H(2, 4)$	$H(2, 5)$	$H(2, 6)$	$H(2, 7)$	D_2
	3	PEV	$H(3, 1)$	$H(3, 2)$	$H(3, 3)$	$H(3, 4)$	$H(3, 5)$	$H(3, 6)$	$H(3, 7)$	D_3
	4	AC (hysteresis)	$H(4, 1)$	$H(4, 2)$	$H(4, 3)$	$H(4, 4)$	$H(4, 5)$	$H(4, 6)$	$H(4, 7)$	D_4
	5	AC (inverter)	$H(5, 1)$	$H(5, 2)$	$H(5, 3)$	$H(5, 4)$	$H(5, 5)$	$H(5, 6)$	$H(5, 7)$	D_5
	6	Dish washer	$H(6, 1)$	$H(6, 2)$	$H(6, 3)$	$H(6, 4)$	$H(6, 5)$	$H(6, 6)$	$H(6, 7)$	D_6
	7	Clothes washer	$H(7, 1)$	$H(7, 2)$	$H(7, 3)$	$H(7, 4)$	$H(7, 5)$	$H(7, 6)$	$H(7, 7)$	D_7
	8	Dryer	$H(8, 1)$	$H(8, 2)$	$H(8, 3)$	$H(8, 4)$	$H(8, 5)$	$H(8, 6)$	$H(8, 7)$	D_8
2	9	Water heater	…							…
	10	Swimming pool	…							…
	11	PEV	…							…
	12	AC (hysteresis)	…		$H(i, j)$					D_i
	13	AC (inverter)	…							…
	14	Dish washer	…							…
	15	Clothes washer	…							…
	16	Dryer	…							…

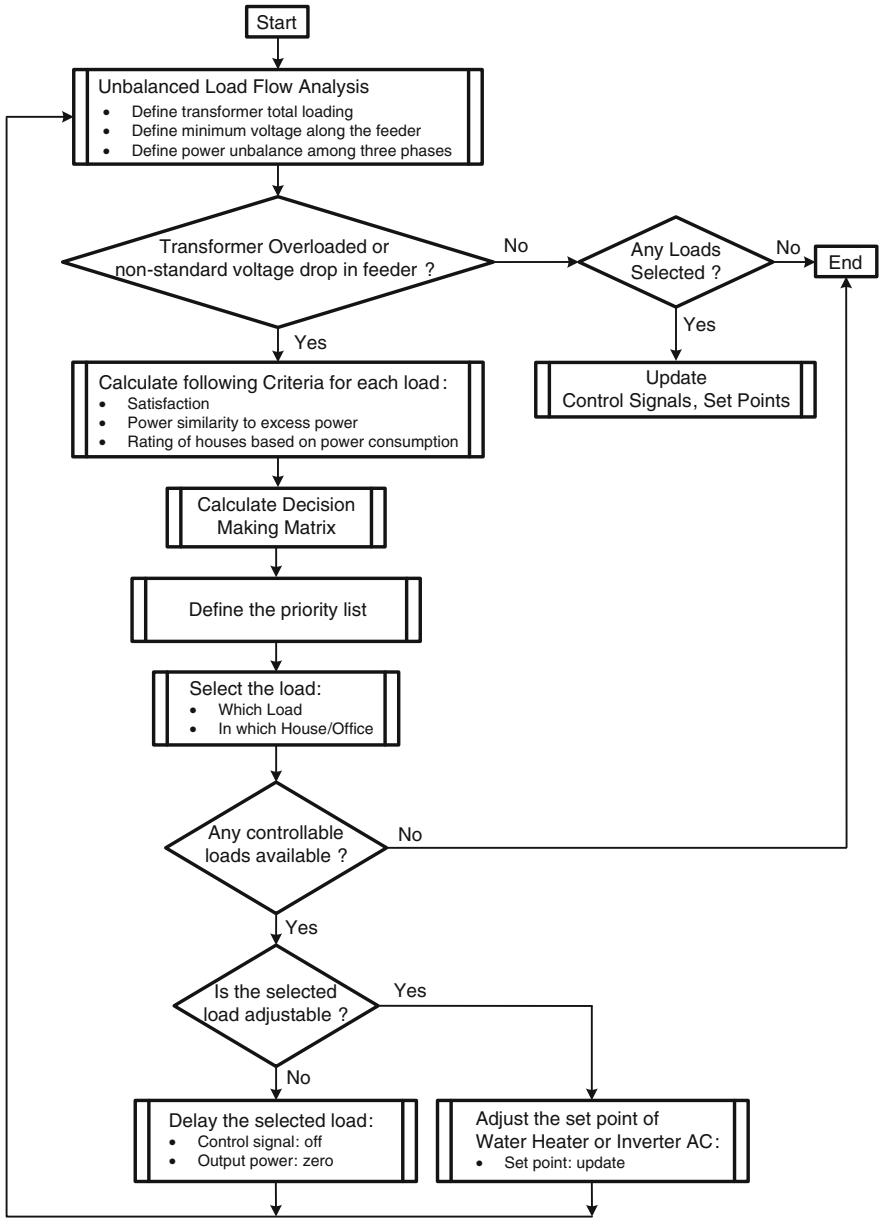


Fig. 3 Flowchart of the proposed smart demand management system including MODM

It is to be noted that the maximum load allowable for each customer or order of load control can also depend on some payment mechanism. If such a tariff adjustment is available in the network, it will be easy to be applied to the proposed MODM method by adjusting new weighting values for the criteria. However, in this study, the electricity price criterion is not considered for load selection.

4.1 Defining Decision Making Matrix

Each of the above mentioned criteria have a different weighting. These weightings (in range [0 1]) are defined based on which criterion is more important than others when selecting a load to be delayed or adjusted. For prioritizing the loads to be controlled, a decision making matrix is used. For this purpose, a numerical value in range [0 1] is allocated for all alternatives (each controllable device in the network) based on the different criteria. All these data are set in a matrix as shown in Table 1.

In Table 1, B_j ($1 \leq j \leq 7$) represents the weighting values for jth criterion. This weighting represents the importance of each criterion for load control in comparison with the other criteria. In this table, $H(i, j)$ is the rank of ith alternative among other alternatives in the same house from the jth criterion point of view where $0 \leq H(i, j) \leq 1$. Therefore, for ith house and a specific jth criterion, $H(i, j)$ is calculated. Then values of $H(i, j)$ are calculated for other criteria in that house. Similarly, $H(i, j)$ is calculated for all controllable devices in the other houses. Finally D_i, the control priority value for ith criterion, is calculated independently for each alternative in the network as

$$D_i = \sum_{j=1}^{7} H(i,j) \times B_j \tag{1}$$

The alternative with highest D_i is chosen as the first alternative for control. The flowchart of the control scheme is shown in Fig. 3.

5 Network Modeling and Analysis Methodology

A MATLAB-based simulation was developed to simulate the low voltage distribution network with an arbitrary number of houses, each containing a number of the above mentioned electric appliances with different time usage patterns. Since most appliances turn ON, run for a certain time at constant power/impedance and then turn OFF, the simulation was made event-based. The simulation comprised a main routine plus a number of appliance modules which simulate the power

characteristics of all appliances of that type. In addition, a calculation is made of the timing of next switching event of that appliance type, and that appliance module is run at that time. The active and reactive power consumptions (or load impedances) are maintained constant between switching events. The simulation also has a fixed time step clock which generates regular events at a user defined interval (typically 5 min). Note that at each event, only the appliance module which contains the next switching device is updated since the power of other appliances is not changed at that time instant. This makes the simulation computationally efficient [29]. The flowchart of the developed program is illustrated in Fig. 4.

An unbalanced load flow analysis is considered and integrated into the developed model. The load flow calculates the voltage along the feeder and the network power loss which are used later for calculating the transformer total loading. It also calculates the power unbalance among the three phases of the system. The load flow runs after each load change continuously in the developed program.

The residential load modeling and unbalanced load flow analysis for the network is described in detail in the rest of this section.

5.1 Modeling of Residential Type Loads

A detailed system model is utilized in this study. The load models were built up, in MATLAB, from individual device/appliance models which later aggregated to form a house and then a residential feeder. Seventeen different appliance types were modeled in total for each house. Each device was allocated with a power rating and power factor as well as a time usage pattern. The time usage pattern of many appliances is linked to seasonal sunrise and sunset times as well as temperature. Therefore a simple "climate model" was constructed to vary the temperature throughout each day.

Each house was assigned a floor area according to a Gaussian distribution with mean 240 m^2 and standard deviation 20 m^2 corresponding to the Australian 2008 new house data [30]. The floor area is used as a reference in the calculation of number and rating of some appliances.

The power ratings of appliances were generally determined using mean values based on the ratings of common models in Australian market while usage patterns were typically derived from an informal assessment of typical behavior patterns. The detailed models of the main devices/appliances, including all parameters used in the study, are discussed below:

1. *Lighting*

The total lighting load of each house was assigned according to a Gaussian distribution with the mean proportional to the house floor area as

$$P_{li,i} = \frac{A_i}{240} \cdot \mathrm{N}(322, 20) \tag{2}$$

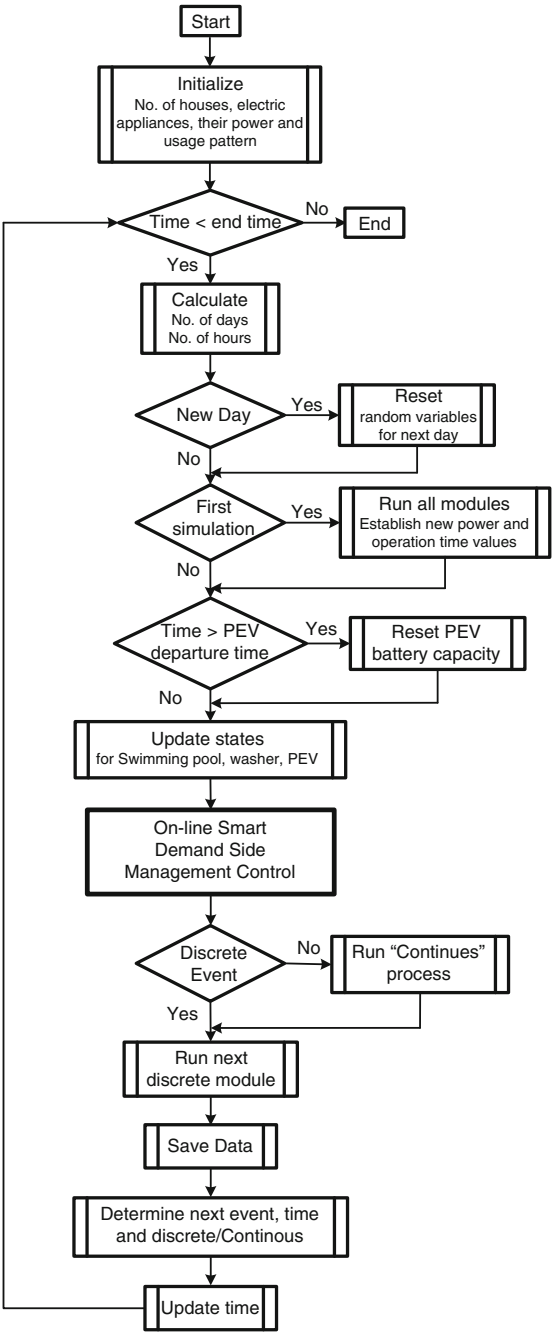


Fig. 4 Flowchart of the MATLAB–based developed analysis program

Table 2 Lighting load turn-on and turn-off times

	Turn-on time TON	Turn-off time TOFF
AM	N(06:00, 1:00) or no turn-on if $T_{sunrise} > T_{ON}$	Earliest of TON + N(02:00, 0:20) and $T_{sunrise}$
PM	Latest of N(18:00, 0:30) and T_{sunset}	N(23:00, 01:00)

where $P_{li,i}$ is the lighting load in kW of house i and A_i is the floor area of house i in m^2. The function N(x,y) is the normal distribution with an average of x and variance of y. The power factor is assumed to be unity. The mean morning and night lighting loads were calculated as 50 and 80 % of the total lighting load. The morning and evening turn-on and turn-off times were based on informal observations as well as accounting for sunrise and sunset times (which vary seasonally). The turn-on and turn-off times were determined as in Table 2.

2. *Fridges and Freezers*

In reality, refrigerators and freezers operate at ON/OFF duty cycles determined by hysteresis temperature controllers. In this study, in order to limit the computational complexity, the thermal dynamic model is neglected and both fridges and freezers are modeled as loads operating at ambient temperature dependant duty cycles. Each house was assumed to have at least one fridge. The probability of second fridge was related to house floor area as:

$$\begin{aligned} P(fr2,i) &= \min(0.1 + (A_i - 240)/240, 1) \\ N_{fr,i} &= 1 + (\text{rand} < P(fr2,i)) \end{aligned} \tag{3}$$

where min(.) is the minimum function, rand(.) is a uniformly distributed random variable and rand $< P(fr2,\ i) = 1$ if true or 0 if false. The fridge period and duty cycle were

$$\begin{aligned} Per_{fr,i} &= \text{N}(0:50, 0:05) \\ D_{fr,i} &= \text{N}(0.18, 0.02).T_{amb}/298 \end{aligned} \tag{4}$$

where $Per_{fr,i}$ is the period in minutes, $D_{fr,i}$ is the duty cycle and T_{amb} is the ambient temperature in Kelvin. The turn-on time of the fridge within the period is random.

The probability of each house having a separate freezer was similarly related to the floor area. In addition, the freezers duty cycle and period were similarly modeled with the mean period of 60 min and the mean duty cycle of 0.16. The power factor was assumed to be 0.85 and the power ratings of the fridges and freezers (in kW) were determined by

$$\begin{aligned} P_{fr,i} &= \text{N}(0.47, 0.04) \\ P_{fz,i} &= \text{N}(0.4, 0.04) \end{aligned} \tag{5}$$

Table 3 Cooking appliance power ratings and times of use

	Stovetop cooker	Conventional oven	Microwave oven
Power rating	N(3.0, 0.3)	N(2.4, 0.2)	N(1.4, 0.14)
T_{ON} (AM)	N(07:00, 0:30)	N(07:00, 0:30) Probability = 0.1	N(07:00, 0:30)
T_{OFF} (AM)	T_{ON} (AM) + N(0:30, 0:03)	T_{ON} (AM) + N(0:30, 0:03)	T_{ON}(AM) + N(0:05, 0:015)
T_{ON} (PM)	N(18:30, 0:30)	N(18:30, 0:30)	N(18:30, 1:00)
T_{OFF} (PM)	T_{ON} (PM) + N(01:00, 0:15)	T_{ON}(PM) + N(0:45, 0:10)	T_{ON} (AM) + N(0:15, 0:02)

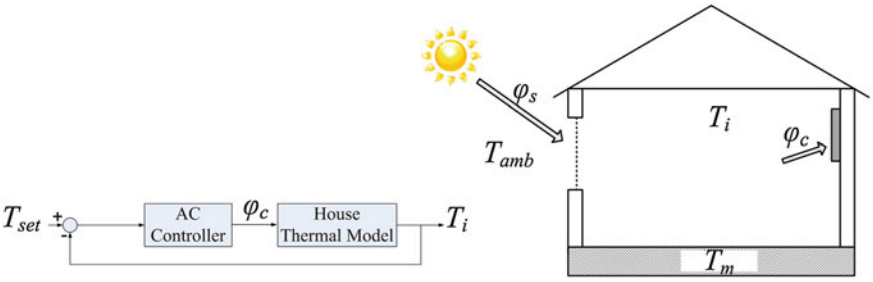


Fig. 5 Indoor thermal modeling and temperature closed-loop control

3. *Cooking appliances*

Stove-top cookers, conventional ovens and microwave ovens were modeled separately. The probability of any house having an electrical stove top cooker and oven was taken as 70 and 90 % respectively (assuming 30 and 10 % gas stove top and oven market penetrations respectively). Each house was assumed to have a microwave oven. The Table 3 shows the calculation of the power ratings and times of use.

4. *Air Conditioners (ACs)*

ACs can have a major impact on residential peak load, so they were modeled in detail. All ACs are modeled as closed-loop temperature controlled devices (either hysteresis or inverter-type), attempting to force internal house temperature to follow a set point. The second order dynamic thermal model of a house, presented in [31], is developed. This model takes into account the effect of AC output power, sun radiation, ambient temperature, door and window size, and floor area of the house. This is shown in Fig. 5 and calculated in (6). The parameters of the indoor thermal model for houses used in this study are given in Table 4.

$$\begin{bmatrix} \frac{dT_m}{dt} \\ \frac{dT_{int}}{dt} \end{bmatrix} = \begin{bmatrix} \frac{-1}{r_{int}C_m} & \frac{1}{r_{int}C_m} \\ \frac{1}{r_{int}C_{int}} & \frac{-1}{r_aC_{int}} - \frac{1}{r_{int}C_m} \end{bmatrix} \begin{bmatrix} T_m \\ T_{int} \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ \frac{1}{r_aC_{int}} & \frac{1-LF}{C_{int}} & \frac{A_w}{C_{int}} \end{bmatrix} \begin{bmatrix} T_{amb} \\ \phi_c \\ \phi_s \end{bmatrix} \quad (6)$$

Table 4 Indoor thermal modeling parameters

$C_{m,\ i}$	N(4, 0.4) · A_i/240
$C_{int,\ i}$	N(1.2, 0.12) · A_i/240
$R_{a,\ i}$	N(29.4, 2.9)
$R_{int,\ i}$	N(0.48, 0.048)
$A_{w,\ i}$	N(2.9, 0.29)
LF_i	0 %

where

r_a Outside ambient/Internal air thermal resistance [°C/kW]
r_{int} Internal air/House slab thermal resistance [°C/kW]
C_{int} House internal air thermal capacity [kWh/°C]
C_m House slab thermal capacity [kWh/°C]
φ_c air conditioning cooling [kW]
φ_s solar radiation [kW/m^2]
T_m House slab temperature [°C]
T_{int} Internal temperature [°C]
T_{amb} Outside Ambient temperature [°C]
T_{set} Air conditioner set point [°C]
A_w Effective window area [m^2]
LF loss factor due to door opening and closing [%].

In this study, a residential market penetration of 70 % was assumed, split between traditional hysteresis ACs and inverter ACs. A small percentage of the inverter ACs was assumed to represent ducted air conditioning (air conditioning the entire dwelling). The assumed split between different types of AC is presented in Table 5. Temperatures T_0, T_{100} are expressed in Kelvin.

5. *Clothes Washers, Dryers and Dishwashers*
All houses were assumed to have clothes washers, with 30 % assumed to have clothes dryers and 50 % dishwashers. Clothes dryer operation was assumed dependent on weather (i.e. a rain parameter) and operated only once clothes washing were completed. All washers and dryers are modeled as constant power loads; the variation of power through different washing and/or drying cycles was not modeled. The parameters of washers/dryers used in this study are as in Table 6.

6. *Electric Water Heaters*
Electric water heaters are still common in many parts of the world including in Australia. In this study, a market penetration of 85 % was assumed. The thermal dynamics of the water heater are modeled and it is assumed that they are not operating under a pre-existing direct load control system. Electric water heaters

Table 5 AC parameters

	Hysteresis AC	Inverter AC	Ducted AC
No of Houses	Total 15 % houses: –10 % in a single room –5 % in 2 rooms	Total 50 % houses: –30 % in a single room –20 % in 2 rooms	Total 5 % house
Rating	N(2.0, 0.2) per room	N(2.3, 0.2) per room	N(5.2, 0.2)
Cycle	N(0:30, 0:05)	N/A	N/A

Table 6 Washer/dryer operating parameters

	Clothes washer	Clothes dryer	Dish washer
Rating	N(1.0, 0.1)	N(2.8, 0.3)	N(2.0, 0.2)
T_{ON}	N(8:30, 0:30) or N(20:00, 0:30)	Clothes washer T_{OFF} + N(0:15, 0:01)	N(07:30, 0:30) or N(20:30, 1:00)
T_{OFF}	T_{ON} + N(01:00, 0:10)	T_{ON} + N(01:00, 0:10)	T_{ON} + N(01:00, 0:02)

were modeled as dependent on ambient temperature and water consumption rate. The parameters of water heaters used in the study are as in Table 7.

7. *Swimming Pool Pumps*

Swimming pool pumps represent large loads that operate for long period of time (e.g. 1.5 kW for 6–8 h per day). In this study a 50 % penetration of swimming pool pumps was assumed. The pool pumps were assumed to operate each day with the total operation time varying from a mean of 6 h in winter to a mean of 8 h in summer. The operation was assumed to be split into two equal periods of operation occurring randomly during the day. The parameters of swimming pool pumps used in this study are as in Table 8.

8. *Plug-in Electric Vehicles (PEVs)*

PEVs are widely thought to be a significant future load for residential power systems, especially when high clusters of PEV owners occur in a certain area. In this study, a total of 25 % penetration of PEVs were considered with a mean battery capacity of 20 kWh (corresponding to the average of the Chevy Volt and Nissan Leaf battery capacities) [31]. It was assumed that the vehicles travel an average of 50 km per day and the batteries charge at a constant rate of 15 A at 230 V until either fully charged or the householder departs. For uncontrolled charging, the PEVs were assumed to start charging at a constant rate from when the customer arrives home until the battery is fully charged or the customer departs (whichever occurs first). Average charge/discharge efficiency of 85 % was considered. The parameters of PEVs used in this study are as in Table 9.

Table 7 Electric water heater parameters

Power rating	N(3.6, 0.3)
Slope of temperature dependency	N(−0.0125, 0.001)

Table 8 Swimming pool pump parameters

Power rating	N(1.5, 0.1)
T_{ON} (AM)	12:00 × rand
T_{ON} (PM)	12:00 + 12:00 × rand
T_{OFF}	T_{ON} + N(8 − 2 × sin(2π × d/730), 0.8), where *d* is the number of days into the year

Table 9 PEV parameters

Charging power rating	230 V, 15 A, unity power factor = 3.45 kW
Daily required charge	N(11.8, 2 kWh)
$T_{arrival}$	N(17:00, 1:00)
$T_{departure}$	N(07:30, 0:30)

9. *Televisions and Personal Computers*

The number of Televisions (TVs) and Personal Computers (PCs) per household was assumed to be in relation the house floor area. The operational times are assumed to be a small amount in the morning and much larger durations in the afternoon and evening to model both adult and children occupants. The parameters of TVs and PCs used in this study are as in Table 10 where round(.) is the rounding function to the nearest integer and ceil(.) is the rounding up function to the next highest integer.

10. *Standby Power and Miscellaneous Appliances*

A small random constant power was allocated to each house to account for standby power and miscellaneous continuously operating appliances (e.g. broadband modems). This amount was changed each day randomly.

11. *Period of No Occupancy*

During week days, there is a high probability that many of the houses are unoccupied for significant periods of the day. During these unoccupied periods, it is highly unlikely that ACs, lighting, cooking appliances, TVs and PCs are operated (albeit that occasionally some people may leave some loads on). To simulate this effect 90 % of the houses were allocated a stochastic "no occupant" period during each week day where operation of the aforementioned appliances was excluded.

Table 10 TV and PC parameters

	TVs	PCs
No. devices per house	Round(N(1.5, 0.3) × Ai/240)	Ceil(N(1.5, 0.3) × Ai/240)
Power rating per device	N(0.18, 0.02)	N(0.26, 0.02)
T_{ON} (AM)	N(07:00, 0:30) Probability = 0.5	N(07:00, 0:30) Probability = 0.5
T_{OFF} (AM)	T_{ON} (AM) + N(0:30, 0:03)	T_{ON} (AM) + N(0:30, 0:03)
T_{ON} (PM)	N(16:00, 2:00)	N(16:00, 2:00)
T_{OFF} (PM)	T_{ON} (PM) + N(2:00, 0:30)	T_{ON} (AM) + N(2:30, 0:30)

5.2 *Unbalanced Load Flow Analysis*

An unbalanced sweep forward–backward load flow [32] is considered and integrated into the developed model. The load flow calculates the voltage along the feeder and the network power loss which are used later for calculating the transformer total loading. It also calculates the power unbalance among the three phases of the system.

For this, first, modified Carson's equations [32] are utilized for calculation of self and mutual impedance of the conductors in the 50 Hz system as

$$Z_{ii} = r_i + 0.04934 + j0.062832\left(Ln\frac{1}{GMR_i} + 7.10988\right) \tag{7}$$

$$Z_{ij} = 0.04934 + j0.062832\left(Ln\frac{1}{D_{ij}} + 7.10988\right) \tag{8}$$

where i and j is the phase conductor (i.e. A, B, C or Neutral), Z_{ii} is the self-impedance of conductor i (in Ω/km), Z_{ij} is the mutual impedance between two conductors i and j (in Ω/km), r_i is the AC resistance of conductor i (in Ω/km), GMR_i is the Geometric Mean Radius of conductor i (in cm) and D_{ij} is the distance between conductor i and j (in cm). Hence, the non-transposed characteristics of the conductors, image conductors below ground and network configuration are considered in the studies. Figure 6a shows the considered line configuration in this study [33]. The three-phase four-wire line segment between two adjacent buses of $k-1$ and k is also shown in Fig. 6b. From (7) and (8), the equivalent impedance for this line section is expressed as

$$[Z_{abcn}] = \begin{bmatrix} Z_{aa} & Z_{ab} & Z_{ac} & Z_{an} \\ Z_{ba} & Z_{bb} & Z_{bc} & Z_{bn} \\ Z_{ca} & Z_{cb} & Z_{cc} & Z_{cn} \\ Z_{na} & Z_{nb} & Z_{nc} & Z_{nn} \end{bmatrix} \tag{9}$$

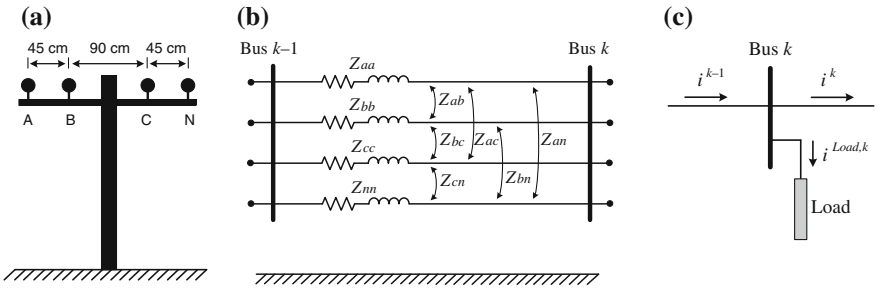


Fig. 6 **a** Low voltage feeder configuration, **b** impedance equivalent of a line segment between two buses, **c** PQ bus model

Assuming the transformer with a delta/star-grounded connection, which is the common distribution transformed in Australia, and using Kron reduction, (9) can be rewritten as

$$[Z_{abc}] = \begin{bmatrix} Z_{aa-n} & Z_{ab-n} & Z_{ac-n} \\ Z_{ba-n} & Z_{bb-n} & Z_{bc-n} \\ Z_{ca-n} & Z_{cb-n} & Z_{cc-n} \end{bmatrix} \tag{10}$$

For simplicity, all calculations are carried out in per unit. Starting with a set of initial values (e.g. flat voltage set), the load currents are calculated as

$$\left[I_{abc}^{Load,k}\right] = conj\left(\frac{\left[P_{abc}^{Load,k}\right] + j\left[Q_{abc}^{Load,k}\right]}{\left[V_{abc}^{k}\right]}\right) \tag{11}$$

where $[I_{abc}^{Load,k}]$ is a vector of three-phase load current connected to bus k, $[V_{abc}^{k}]$ is a vector of three-phase voltage of bus k, $[P_{abc}^{Load,k}]$ and $[Q_{abc}^{Load,k}]$ are respectively a vector of three-phase active and reactive power consumption of the residential load connected at bus k.

The sum of the all load currents will flow from the first bus (transformer secondary side) to the second bus. Therefore, as shown in Fig. 6c, the current between two adjacent buses is

$$[I_{abc}^{k}] = [I_{abc}^{k-1}] - [I_{abc}^{Load,k}] \tag{12}$$

Hence, the voltage of bus k can be calculated based on the voltage of bus $k-1$ in its upstream and the current passing between two buses as

$$[V_{abc}^{k}] = [V_{abc}^{k-1}] - [Z_{abc}][I_{abc}^{k}] \tag{13}$$

Once the voltage at bus k is calculated, the load current in that bus will be updated from (11) and then using (12) the current flowing from bus k to $k+1$ in its downstream are updated.

Similar to the line segment, the equivalent impedance of the delta/star-grounded distribution transformer between its primary and secondary buses is expressed as

$$[Z^k_{abc}] = z_t \times I \tag{14}$$

where z_t is the phase impedance of the transformer and I is the identity matrix. Now, the secondary-side voltage of the transformer are calculated from its primary-side voltage as [32].

$$[Vt^S_{abc}] = [A][Vt^P_{abc}] - [Zt_{abc}][I_{abc}] \tag{15}$$

where $[Vt^P_{abc}]$ and $[Vt^S_{abc}]$ are respectively the primary and secondary-side phase voltages of the transformer and $[I_{abc}]$ is a vector of three-phase current passing through the transformer and

$$[A] = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 & 0 & -1 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \tag{16}$$

The flowchart of the unbalanced load flow analysis is illustrated in Fig. 7.

6 Simulation Results

For verifying the efficacy of the proposed control a few sample results are given below. First, let us consider an 11 kV feeder supplying in total 20 residential low voltage feeders. Each residential feeder is supplied through a 100 kVA 11 kV/400 V pole-mounted transformer. Let us consider 20 houses in each low voltage feeder (i.e. ADMD of 5 kVA). The length of each low voltage feeder is assumed to be 400 m where 10 poles are installed at 40 m distances from each other. The simulation is conducted over a 48-h week-day summer period (max ambient temperature of 35 °C at 2 pm). In addition, a 40 % penetration level of rooftop PVs, all equipped with batteries, are considered. Each distribution transformer is equipped with a controller which is responsible for the loads in its downstream.

In addition, the flexibilities assumed for each controllable device in this study are listed in Table 11. Note that, these values are used just as an assumption and can be changed depending on stakeholders' feedback. In Table 11, each device is given a number between one and eight which represents that device in the simulation results figures. Note that the final device selection in each house depends on the priority (randomly generated), flexibility and criteria weightings. The assumed weighting of the MODM criteria in this study is as shown in Table 12.

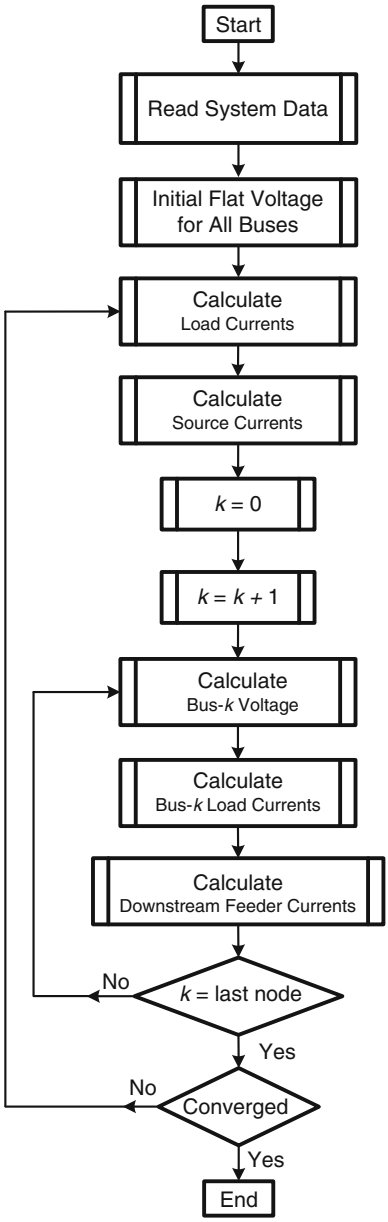


Fig. 7 Flowchart of developed unbalanced load flow analysis

Table 11 Controllable device number allocation and flexibility

Controllable device	Water heater	Pool pump	AC (hysteresis)	AC (inverter)	PEV	Dish washer	Clothes washer	Dryer
Number	1	2	3	4	5	6	7	8
Flexibility	0.5	0.9	0.7	0.7	0.9	0.2	0.2	0.2

Table 12 Weightings of MODM criteria

Priority	Flexibility	Satisfaction	Power matching	High power	Power balance	Battery
0.8	0.5	0.7	0.7	0.9	0.8	0.6

First, let us assume the smart demand management system controls the loads only if the distribution transformer is overloaded and it does not consider voltage drop and power unbalance in selecting the loads.

Figure 8a shows the total aggregated apparent power supplied by a 100 kVA residential feeder transformer in the presence of PEVs. The specific load shape will vary somewhat stochastically from day to day and from area to area. As seen from this figure, although the average apparent power supplied by the transformer is around 60 kVA per day, it reaches a peak of 150 kVA. The transformer is overloaded by 50 % for a 2-h period per day.

Let us now assume that the proposed control system is applied to the network. The total aggregated apparent power supplied by the residential distribution transformer is shown in Fig. 8b. As seen in this figure, the peak load power is limited to 100 kVA hence verifying the efficacy of the proposed control system in controlling load. Comparing Fig. 8a and b, a load increase is seen after the control operation period which is the result of the delay/adjustment of specific loads by the controller.

As seen in Fig. 8b, the load varies "instantaneously" around the transformer rating. This is due to the fact that the control action occurs in 2-min time intervals whereas load changes can occur at any time. Therefore, after a control action at t_1, some load might increase (or turn on) during the 2 min time interval which will cause the load to exceed the rating of the transformer until the control action at $t_1 + 2$ min.

Now, let us consider a 2 MVA transformer in a substation supplying 20 low voltage feeders, each with a rating of 100 kVA. It is assumed that the discussed smart demand management is developed for all 20 low voltage feeders in this network. Figure 8c shows the total apparent power of the 2 MVA upstream transformer in the substation in 48-h period. As it can be seen from this figure, in the uncontrolled case, the substation transformers experience an overload around 30 % for a period of 3-h. However, if the proposed control method is applied for all the residential loads which are supplied by each distribution transformer, the peak load is limited to 2 MVA as shown in this figure. This demonstrates the

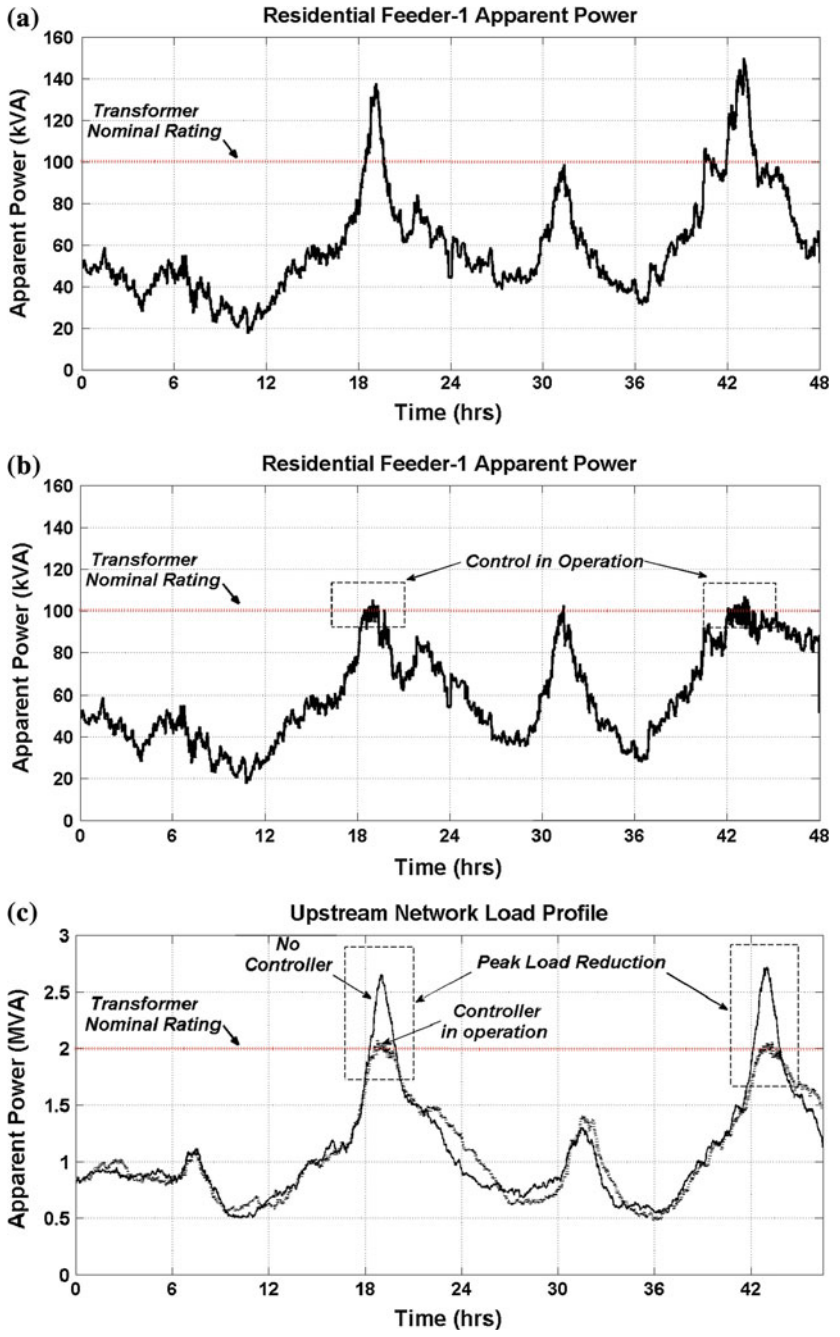


Fig. 8 Total apparent power of: **a** the residential distribution transformer without any control. **b** The residential distribution transformer with demand management. **c** The substation feeding 100 distribution transformers in an area with and without the demand management system

potential for savings in the upstream network electrical infrastructure. The difference in load profiles is a result of some peak loads being transferred to off peak periods.

It is important for the control scheme to have a minimal impact on consumers. To verify this, sample operation states of PEVs and swimming pools as (delayed) controllable loads are given below. Figure 9a shows the ramp waveforms indication the battery charging state of a few PEVs of the network. It is to be noted that each color line in this figure represents the charging status of a sample PEV within the network. As this figure shows they are being charged from an arbitrary initial value to full (100 %) before their departure time. This value is reset to an arbitrary value at 12 pm everyday and they start being charged at an arbitrary time in the evening when the cars return home. The flat sections of the traces show that they are disconnected by the controller at some stages and again reconnected, nevertheless all vehicles complete charging. This figure verifies that even though the controller has delayed the charging for some of them, they have all been fully charged before departure time.

Figure 9b shows the swimming pool pump remaining operation time for a few houses in the network, which is reset every day. It is to be noted that each color line in this figure represents a sample swimming pool within the network. As seen in this figure, each swimming pool has a randomized required operational time per day (around 7–9 h) and they have all completed operation (i.e. they fall to zero) within each 24 h period. All other controllable devices have the same operational characteristic.

Inverter type ACs and water heaters are large loads in the network and have adjustable characteristics based on changing the set points of room temperature and tank water temperature. Based on assumption in this study, inverter ACs will operate when the internal house temperature exceeds its set point providing it is occupied. If an inverter AC is selected, the transformer controller will increase the temperature set point by 1 °C. Fifteen minutes later this set point will be reset to its original value, thus preventing set point divergence from the householders desired level. In Fig. 9c, the AC inverter set point increase and reset is shown for three sample houses. The set point increase results in less electric power consumption by the AC as shown in this figure.

However, this set point change should not appreciably worsen the customer satisfaction. This is investigated in Fig. 10. In Fig. 10a, the ambient and internal temperature in 48-h period is shown. As it can be seen the internal temperature is kept around 25 °C (set point). In Fig. 10b, the electric power consumption of the AC is shown. It can be seen that when the internal temperature rises beyond the set point limit, the AC turns on and turns off when the internal temperature is reduced to the set point. In Fig. 10c the AC satisfaction is shown. When the people are away, the AC is off so "satisfaction" is decreased but when the people are at home, it is kept close to 100 %. In Fig. 10d, the set point and its variation based on the controller command is shown. It can be seen that around 43:00 h, there are several commands for set point increase.

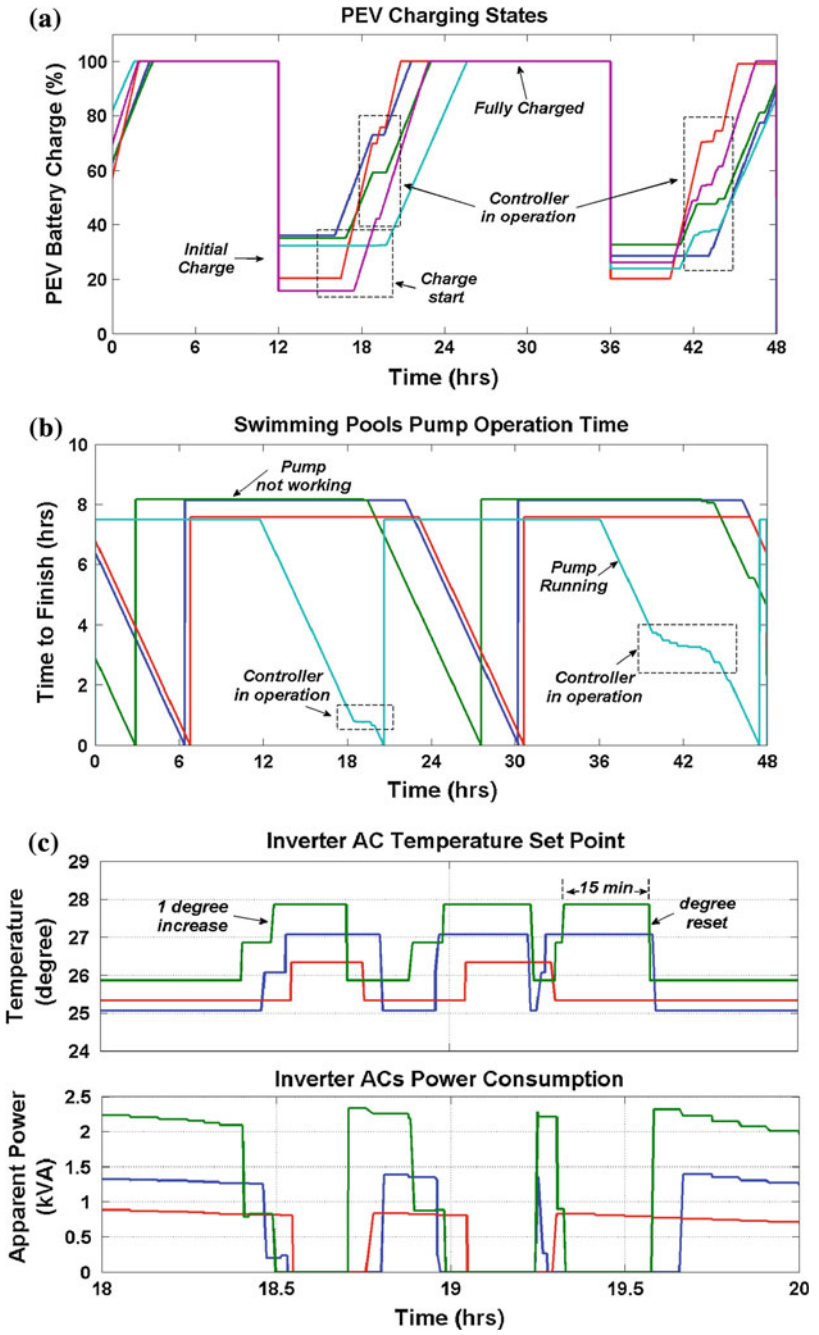


Fig. 9 **a** PEVs battery charging states. **b** Swimming pool pump operation characteristic. **c** Temperature set point and relevant apparent power consumption change for sample residential inverter ACs in network

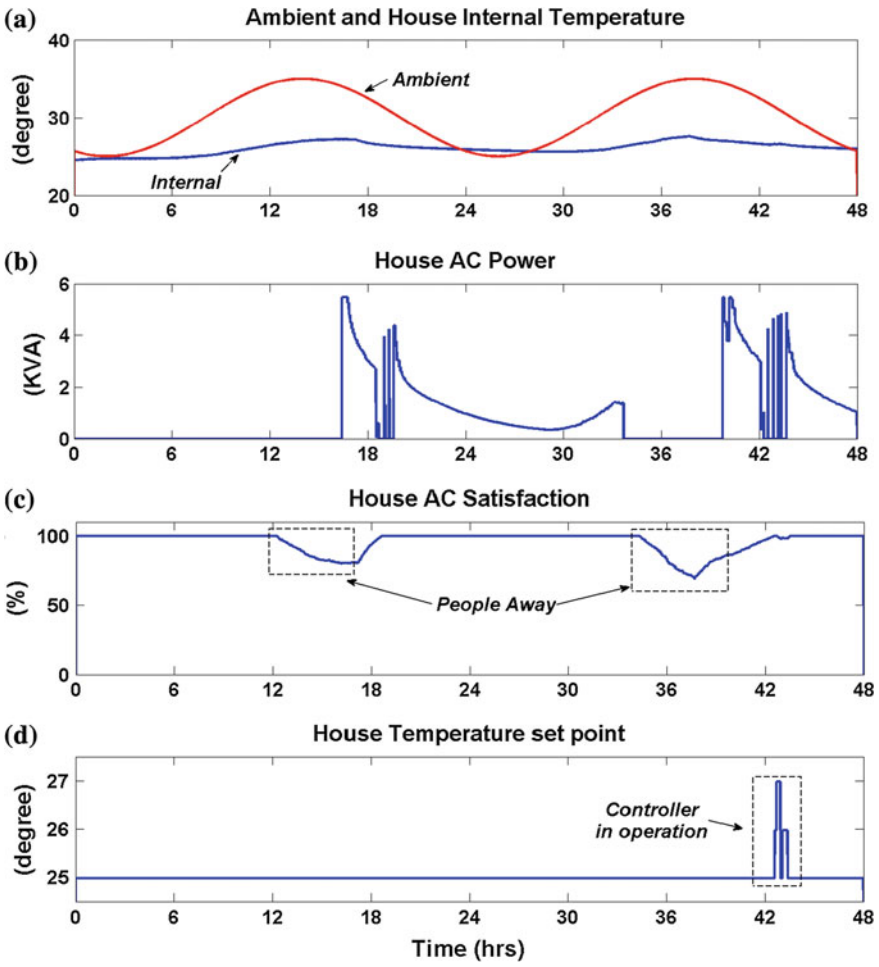


Fig. 10 **a** Ambient and house internal temperature variation, **b** AC electric power consumption, **c** AC satisfaction, **d** AC set point

The same behavior is evident for other ACs and water heaters in the network.

It is of interest to study the number of control commands applied by the transformers' controller. In Fig. 11a, the number of control commands applied over a 48 h period to each controllable device is shown for four sample houses in a residential feeder. The controllable device number was given in Table 3. From this figure, it can be seen that the number of control commands differs from house to house based on the status of any controllable device running in each house, as well as their objectives (power, satisfaction, priority and flexibility).

In Fig. 11b, the total number of the low level control commands applied over a 48 h period for each customer in each feeder individually and for each type of

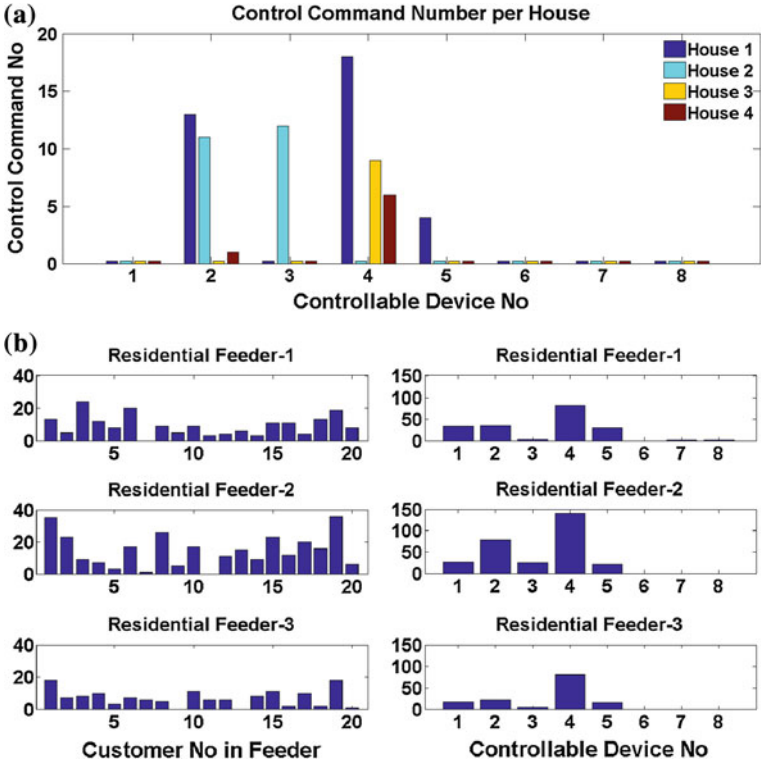


Fig. 11 Number of control commands applied to different controllable devices in: **a** Four sample houses of residential feeder-2. **b** Each customer in a feeder individually (*left*) and for a specific controllable device in each feeder (*right*)

controllable device in that feeder is shown. It can be seen that the number of commands for load control differs from a house to house. It is also shown that inverter type ACs experience the highest number of control actions.

The results presented in Figs. 8–11 were based on the smart demand management system which was only preventing transformer overloading. The voltage drop and power unbalance improvement as well as discharge control of the batteries of rooftop PVs were not considered. In the rest of this section, these criteria are considered in the smart demand management system.

Now let us consider a low voltage distribution feeder with 60 houses supplied through a 300 kVA, 11 kV/400 V pole-mounted distribution transformer. It is assumed 20 houses are connected to each phase. The total apparent power supplied by the distribution transformer is shown in Fig. 12. Note that in Figs. 12–14, *DSM* represents the management system without considering voltage drop and power unbalance while *DSM* and *VR* represents the management system considering those issues.

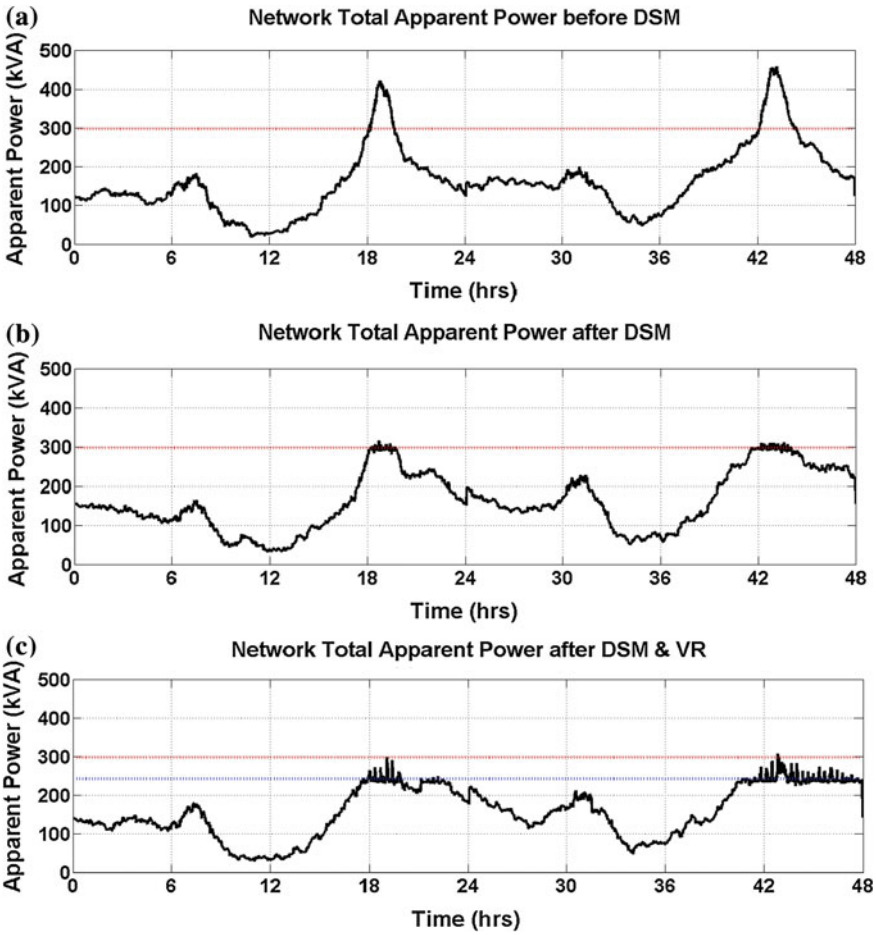


Fig. 12 Network total apparent power in different cases

The network voltage profile is also shown in Fig. 13. As seen from Fig. 12a, although the average apparent power supplied by the transformer is below 200 kVA per day, it reaches a peak of 450 kVA. This leads to excessive voltage drop in the network during network peak hours, as shown in Fig. 13a. By applying the DSM, the apparent power supplied by the transformer can be limited to 300 kVA as shown in Fig. 12b; however, the voltage drop still exceeds required limits as shown in Fig. 13b. By applying the voltage regulation method, as described before, and adjusting the transformer rating to 80 % of its nominal

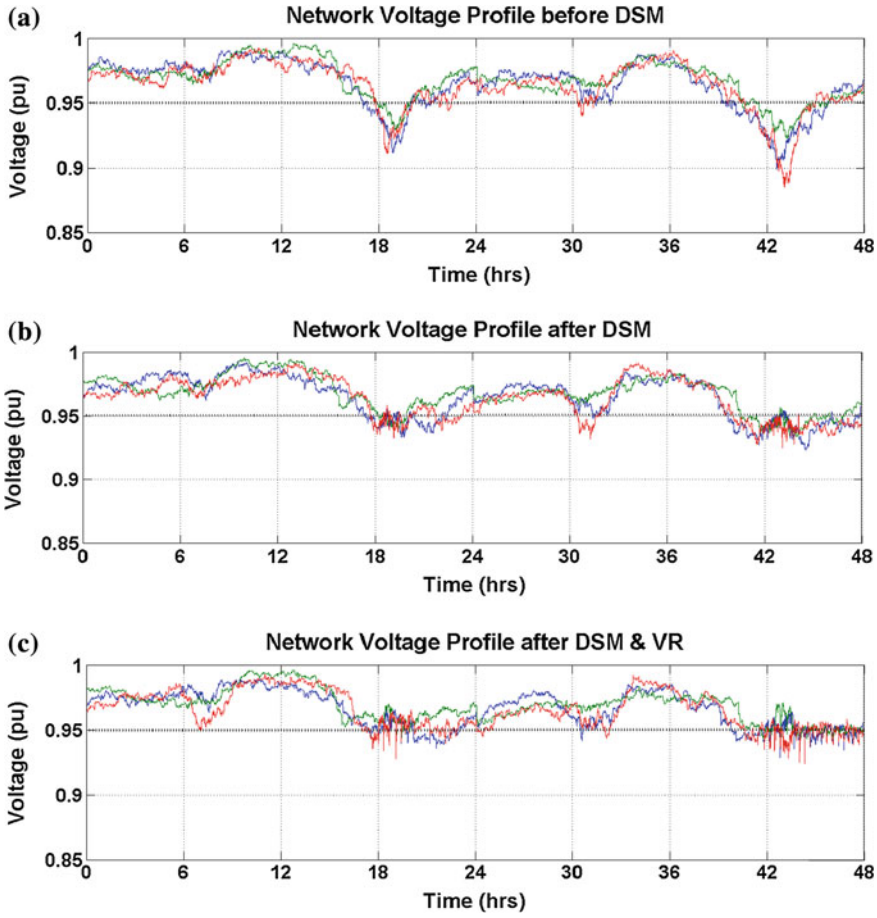


Fig. 13 Network voltage profile in different cases

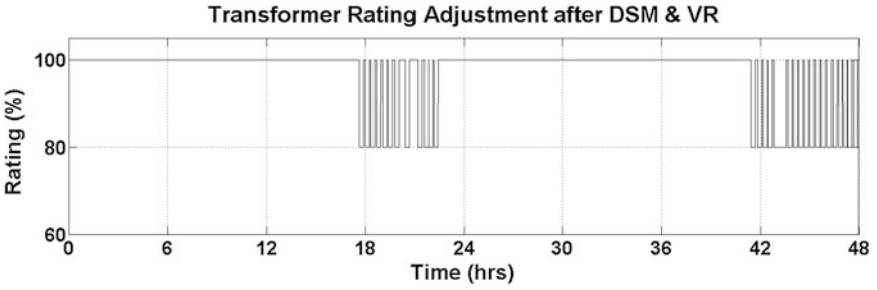


Fig. 14 Transformer rating adjustment during voltage regulation

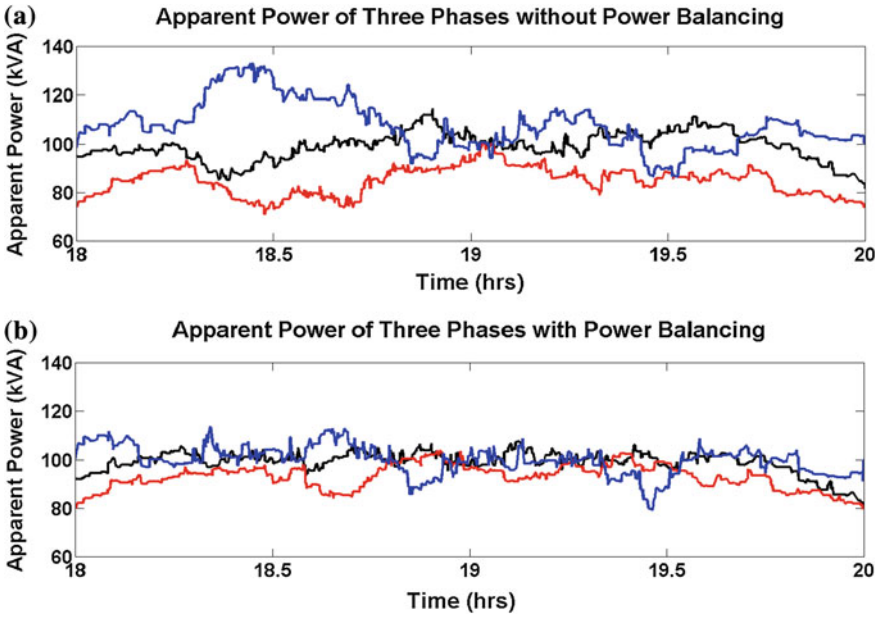


Fig. 15 Comparison of power unbalance criteria in MODM with a weighting of: **a** zero and **b** 80 %

value, the load management and voltage regulation is achieved simultaneously, as shown in Figs. 12c and 13c. The transformer rating adjustment is shown in Fig. 14.

As described before, it is expected that by applying the phase balancing criterion in MODM method, the power in the three phases of the system to be more balanced. Figure 15a shows the apparent power in the three phases of the system when the phase balancing criteria is disabled while Fig. 15b shows the result for the same case while the power balancing criteria is enabled. These results verify the effect of the proposed power balancing criterion.

The power generation output for a sample 2-kW PV system in the network is shown in Fig. 16a. As it can be seen from Fig. 16b, this power is always stored in the installed battery until it is fully charged. Then, the surplus of the PV generation flows into the grid. This energy is later discharged in the network when there is a voltage drop problem in the network as shown in Fig. 16c.

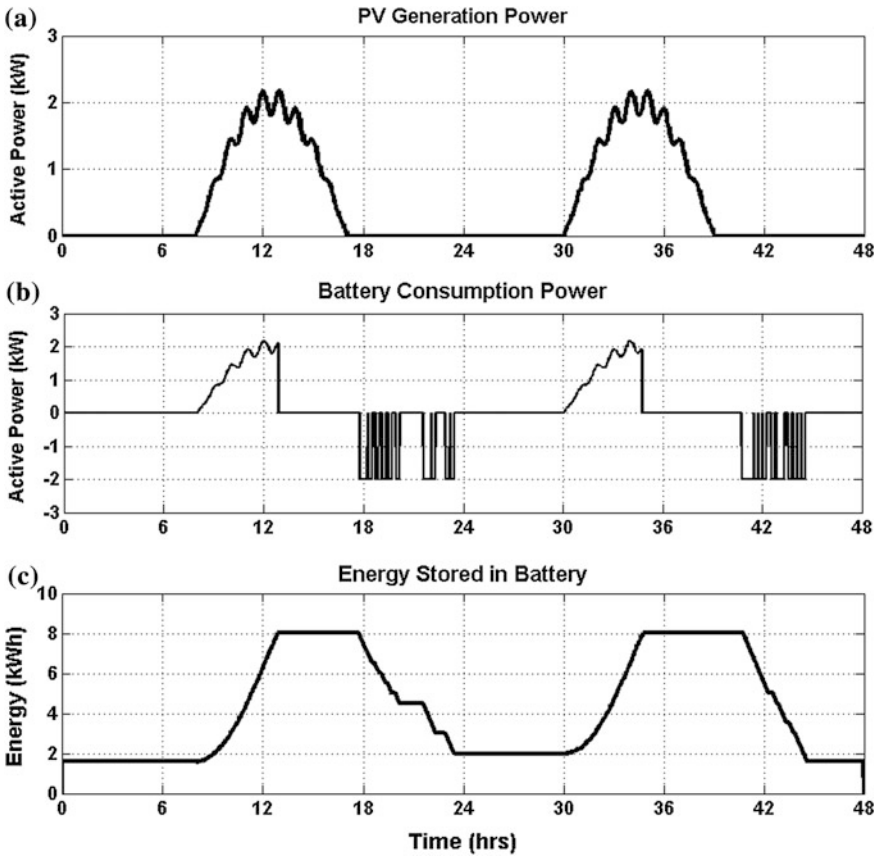


Fig. 16 **a** Power generation curve for a sample 2-kW PV, **b** power consumption curve for a sample 8-kWh battery, **c** energy stored in the battery

7 Conclusion

A novel real time smart direct demand management system was discussed for peak load reduction in low voltage residential distribution networks in this chapter. The network voltage drop was also improved by the coordination of discharging of the installed batteries by the householders. In addition, the system minimized the power unbalance among the three phases of the network for their rooftop PVs. The system utilizes low cost controllers with two-way communications capability, installed in each house and at the supply transformers, to monitor and control the loads. A multi-objective decision making process was proposed to select which load(s) to be delayed or reduced. The algorithm uses several criteria, each with different weightings to reduce the load while minimizing the impact on consumers.

The simulated results showed the capability of the system to control overloads, improve voltage drop and reduce power unbalance among the phases while maintaining the required load objectives.

References

1. H.L. Willis, *Power Distribution Planning Reference Book* (Marcel Dekker, New York, 2004)
2. Energex Annual Report, (2008/2009), http://www.energex.com.au/__data/assets/pdf_file/0011/26669/2008_09_Annual_Report.pdf
3. Ergon Energy Annual Stakeholder Report, (2008/2009), http://www.ergon.com.au/__data/assets/pdf_file/0020/9218/2009–Annual_Stakeholder_Report_08–09.pdf
4. J. Taylor, A. Maitra, M. Alexander, D. Brooks, M. Duvall, Evaluation of the impact of plug-in electric vehicle loading on distribution system operations, in *IEEE Power and Energy Society General Meeting* (2009), pp. 1–6
5. K. Schneider, C. Gerkensmeyer, M.K. Meyer, R. Fletcher, Impact assessment of plug–in hybrid vehicles on pacific northwest distribution systems, in *IEEE Power and Energy Society General Meeting* (2008), pp. 1–6
6. P. Mohseni, R.G. Stevie, Electric vehicles: holy grail or fool's gold, in *IEEE Power and Energy Society General Meeting* (2009), pp. 1–5
7. K. Bhattacharyya, M.L. Crow, A fuzzy based load model for power system direct load control, in *IEEE 4th Conference on Control Applications* (1995), pp. 27–32
8. A. Gabaldon, A. Molina, C. Roldan et al., Assessment and simulation of demand–side management potential in urban power distribution networks, in *IEEE Power Tech Conference*, vol. 4 (2003)
9. A. Amato, M. Calabrese, V. Di Lecce, V. Piuri, An intelligent system for decentralized load management, in *IEEE International Conference on Computational Intelligence for Measurement Systems and Applications* (2006), pp. 70–74
10. B. Ramanathan, V. Vittal, A framework for evaluation of advanced direct load control with minimum disruption. IEEE Trans. Power Syst. **23**(4), 1681–1688 (2008)
11. H.Y. Yih, S.C. Ching, Dispatch of direct load control using dynamic programming. IEEE Trans. Power Syst. **6**(3), 1056–1061 (1991)
12. I. Richardson, M. Thomson, D. Infield, A. Delahunty, A modeling framework for the study of highly distributed power systems and demand side management, in *IEEE International Conference on Sustainable Power Generation and Supply (SUPERGEN)* (2009), pp. 1–6
13. C.Y. Chang, C.J. Wu, C.T. Chang et al., Experiences of direct load control using ripple signals in Taiwan power system, in *4th International Conference on Advances in Power System Control, Operation and Management,* vol. 2 (1997), pp. 591–596
14. B.F. Hastings, Ten years of operating experience with a remote controlled water heater load management system at Detroit Edison. IEEE Trans. Power Apparatus Syst. **PAS–99**(4), 1437–1441 (1980)
15. G. Inc. (2010, 5/9/2011), Load Management, http://www.gridpoint.com/utility–solutions/load-management.aspx
16. Y.C. Min, T.H. Yuan, K.C. Keat et al., e2Home: A lightweight smart home management system, in *IEEE 3rd International Conference on Convergence and Hybrid Information Technology (ICCIT)*, vol. 1 (2008), pp. 82–87
17. Lochiel Park http://www.lochielpark.com.au/lochielpark/energy.htm
18. P. Richardson, D. Flynn, A. Keane, Optimal charging of electric vehicles in low-voltage distribution systems. IEEE Trans. Power Syst. **27**(1), 268–279 (2012)

19. M. Takagi, K. Yamaji, H. Yamamoto, Power system stabilization by charging power management of Plug-in Hybrid Electric Vehicles with LFC signal, in *IEEE Vehicle Power and Propulsion Conference (VPPC)* (2009) pp. 822–826
20. F.C. Schweppe, M.C. Caramanis, R.D. Tabors, R.E. Bohn, *Spot pricing of electricity* (Kluwer Academic Publishers, Boston, 1988)
21. A.K. David, Y.Z. Li, Effect of inter-temporal factors on the real time pricing of electricity. IEEE Trans. Power Syst. **8**(1), 44–52 (1993)
22. K. Clement, E. Haesen, J. Driesen, Coordinated charging of multiple plug-in hybrid electric vehicles in residential distribution grids, in *IEEE Power and Energy Society General Meeting* (2009), pp. 1–7
23. F. Shahnia, R. Majumder, A. Ghosh, G. Ledwich, F. Zare, Voltage imbalance analysis in residential low voltage distribution networks with rooftop PVs. Electr. Power Syst. Res. **81**(9), 1805–1814 (2011)
24. J. Cappelle, J. Vanalme, S. Vispoel et al., Introducing small storage capacity at residential PV installations to prevent overvoltages, in *IEEE International Conference on Smart Grid Communications* (2011), pp. 534–539
25. F. Shahnia, A. Ghosh, G. Ledwich, F. Zare, Voltage correction in low voltage distribution networks with rooftop PVs using custom power devices, in *37th Annual Conference of IEEE Industrial Electronics Society (IECON)* (2011), pp. 991–996
26. U. Habiba, S. Asghar, A survey on multi-criteria decision making approaches, in *IEEE International Conference on Emerging Technologies* (2009), pp. 321–325
27. J.A. Jardini, J.L.P. Brittes, L.C. Magrini et al., Power transformer temperature evaluation for overloading conditions. IEEE Trans. Power Delivery **20**(1), 179–184 (2005)
28. http://www.energex.com.au/sustainability/sustainability-rewards-programs
29. F. Shahnia, M.T. Wishart, A. Ghosh, G. Ledwich, F. Zare, Smart demand side management of low-voltage distribution networks using multi-objective decision making. IET Gener. Transm. Distrib. **6**(10), 968–1000 (2012)
30. Australian Bureau of Statistics, Average floor area of new residential dwellings (2009), http://www.abs.gov.au/ausstats/abs@.nsf/featurearticlesbyCatalogue/8BB3F6B866BC35CECA2578A000153026?OpenDocument
31. H. Madsen, J. Holst, Estimation of continuous-time models for the heat dynamics of a building. Energy Build. **22**, 67–79 (1995)
32. W.H. Kersting, *Distribution System Modeling and Analysis* (CRC Press, Boca Raton, 2012)
33. Distribution Construction Standards Handbook (2007), http://www.westernpower.com.au/networkcontractors/Distribution_Construction_Standards_Handbook.html

Smart Inter-Phase Switching of Residential Loads in Low Voltage Distribution Feeders

Farhad Shahnia, Peter Wolfs and Arindam Ghosh

Abstract In order to dynamically reduce voltage unbalance along a low voltage distribution feeder, a smart residential load transfer system is discussed. In this scheme, residential loads can be transferred from one phase to another to minimize the voltage unbalance along the feeder. Each house is supplied through a static transfer switch and a controller. The master controller, installed at the transformer, observes the power consumption in each house and will determine which house(s) should be transferred from an initially connected phase to another in order to keep the voltage unbalance minimum. The performance of the smart load transfer scheme is demonstrated by simulations.

Keywords Distribution feeder · Load transfer · Static transfer switch · Voltage unbalance · Smart grid

F. Shahnia (✉)
Center of Smart Grid and Sustainable Power Systems, Department of Electrical and Computer Engineering, Curtin University, Perth, Australia
e-mail: farhad.shahnia@curtin.edu.au

P. Wolfs
Power and Energy Center, School of Engineering and Built Environment, University of Central Queensland, Rockhampton, Australia
e-mail: p.wolfs@cqu.edu.au

A. Ghosh
School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane, Australia
e-mail: a.ghosh@qut.edu.au

Z. H. Khan et al. (eds.), *Computational Intelligence for Decision Support in Cyber-Physical Systems*, Studies in Computational Intelligence 540,
DOI: 10.1007/978-981-4585-36-1_11,

1 Introduction

Current unbalance and voltage unbalance are among the main power quality problems in low voltage distribution networks [1]. Australia is similar to Asia, Europe and Africa, in that the low voltage residential feeders are in the form of three-phase, four-wire systems, supplied by Delta/Star-Grounded (Dyn-type), three-phase transformers [2]. Majority of the houses have single-phase power supply but larger homes may have a three-phase connection. The voltage unbalance can be very high in networks with voltage drops close to the allowable limits, especially if the houses are distributed unequally among the three-phases [2]. The growing penetration of rooftop photovoltaic cells (PVs) in low voltage residential feeders has increased the voltage unbalance problem. The rooftop PVs output power is intermittent and the number of rooftop PVs connected to each phase is often unequal. This will significantly affect the voltage unbalance in the feeder depending on penetration level, rating and location of PVs in the feeder [3]. The growing penetration of plug-in electric vehicles will further contribute to unbalance. In [4], it was shown that plug-in electric vehicles might lead to high voltage unbalance in low voltage feeders in both charging and discharging modes.

At construction, the utilities aim to distribute the houses equally among the three phases of distribution feeders [2]. Currently, the electric utilities minimize the unbalance problem in low voltage feeders by manually changing the connection phase of some of the customers. However, the best solution is reducing the unbalance at the unbalance generation point. Different methods are already considered for unbalance reduction in low voltage feeders. In [3], some conventional improvement methods such as feeder cross-section increase or capacitor installation are investigated for this purpose. In [5], the application of power electronics-based custom power devices was considered. It was shown that custom power devices can correct the voltage at the Point of Common Coupling (PCC) to a balanced voltage. Hence, if the PCC voltage is balanced, the current drawn from the upstream network will be balanced and the unbalance will not penetrate to upstream network. In [3], reactive power exchange by rooftop PVs was proposed for balancing the voltage. Although this method is very efficient, it will take time for PV inverter standards to evolve.

In modern distribution networks, sectionalizing switches and normally open tie switches are often used for reconfiguration of the network in medium voltage levels. One of the main benefits of network reconfiguration is reducing the network power loss [6], increasing the penetration level of distributed generation units in the network [7], improving power quality [8] and network restoration of service to larger number of customers following a fault [9]. In [10], the network reconfiguration was carried out by simply changing the phase connection of the three phases in the primary side of the distribution transformer for voltage unbalance and power loss reduction. Therefore, based on the known load pattern for each distribution transformer, the optimum phase balancing was carried out. However, this practice is only carried out once and is not dynamic. In [11], using Static

Transfer Switches (STS), a sensitive load was supplied from two different feeders to prevent voltage sag or swell on the load by quickly transferring the input of the load from one three-phase feeder to another three-phase feeder. A similar network reconfiguration and load transfer scheme, derived from [10, 11], can be applied in low voltage feeders to reduce voltage.

In this chapter, a smart dynamic residential load transfer system is discussed. This system consists of a master controller and several distributed end-user controllers and STSs. In this system, the end-user controller, installed at each house, transmits the power consumption of the house to a master controller, installed at the distribution transformer. The master controller will then analyse the network voltage unbalance and total power consumption in each phase and will determine which house(s) should be transferred from initially connected phase to another phase. Then, it will send the proper command to the end-user controller in the selected house(s). Upon receiving a command from the master controller, the end-user controller will activate the STS to change the phase connection for that house accordingly. The STS has a three-phase input and a single-phase output connection to provide all phase connections to a single-phase residential load. For demonstrating the performance of the smart load transfer system, a comprehensive analysis is carried out in MATLAB to demonstrate the voltage unbalance, maximum and minimum of voltage along the feeder and power mismatch between the three phases of the network under consideration. The study later investigates the participation level of the houses in the smart load transfer system in addition to the effect of location of houses along the feeder. The algorithm is later modified to utilize a genetic algorithm-based optimization to achieve the minimal voltage unbalance and power mismatch in three phases.

2 Smart Load Transfer Scheme

Let us consider Fig. 1 which shows a typical low voltage distribution feeder in a suburban area. The feeder supplies a few tens of houses with single phase connections, distributed unequally in the three phases. It is desired to define the candidate load(s) in the network which by transferring them into other phases, the voltage unbalance is minimised along the feeder. As the load characteristics vary in 24-h period, the load transfer needs to be carried out dynamically throughout the day.

The main objective of the smart load transfer system is to ensure voltage unbalance is minimized along the feeder while the power mismatch in three phases in the secondary side of the distribution transformer is also minimized. The preliminary stage of this scheme is that the utilities are aware of the phase connection of each house in the feeder. If this is not known, the method presented in [12] can be utilized for phase identification for each house. A detailed description on the smart load transfer system, its required equipment and communication is given below:

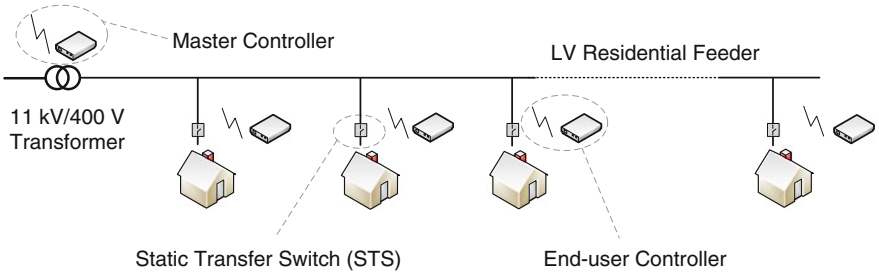


Fig. 1 Schematic diagram of a typical low voltage residential distribution feeder with the discussed smart load transfer system

A. *Smart Meters*

The smart load transfer system requires access to instantaneous power consumption by the houses. For this, all residential participants in the smart load transfer system should be equipped with smart meters [13] that transmit the power consumption of the house to the master controller in 15-min time intervals.

B. *Controllers*

In the smart load transfer system, there are two controllers. The master controller is installed at the distribution transformer and analyzes the network voltage unbalance and power mismatch between the three phases after receiving the power consumption from smart meters in 15-min time intervals. Then, based on its control algorithm, it selects the candidate house(s) which a load transfer is required. The candidate load(s) selection mechanism is described later in this chapter. After selecting the candidate load(s), the controller subsequently sends a control command to the selected house(s).

An end-user controller is installed at each house which is participating in the smart load transfer system. This controller will activate the STS once it receives a control command from the master controller. Then, it will send the confirmation of successful load transfer to the master controller.

The main requirements of the smart load transfer system are:

- Effectiveness: The system should minimize the voltage unbalance along the feeder and the power mismatch in the three phases in secondary side of the distribution transformer.
- Low cost: The system should utilize low-cost hardware.
- Scalable: The system should be easily scalable to larger networks.
- Robustness: The system should be fault tolerant.

C. *Communication*

Different communication methods are already utilized in electric distribution and transmission networks such as Power line carriers, Optical fibre Ethernet, Internet, 3G/4G wireless, WiFi and ZigBee [14]. However, in recent years, ZigBee is the most preferred communication method for data transfer in smart grid applications

in distribution networks. Therefore, in this study, it is assumed that a ZigBee-based communication is available for transferring the control commands from the master controllers to the end-user controllers. The available ZigBee devices along with their range extenders can easily cover an area of 1.6 km and have a data rate up to 250 kHz [15].

Two important parameters for the communication are the cover range and data transfer rate. In the smart load transfer system, the length of the low voltage distribution feeder is assumed to be about 400 m, which is the normal condition in Australia. In addition, the smart load transfer system operates on 15-min time intervals. Therefore, for this application, a very low-bandwidth communication system is sufficient.

D. *Static Transfer Switches*

The load transferring device in the discussed smart load transfer system is a power electronics-based STS, as shown in Fig. 2. The STS is composed of three switching devices, one for each phase. Each switching device is composed of anti-parallel thyristors or a Triac. Overvoltage protection and snubber circuits are in parallel with each Triac [1]. Each Triac is connected to one of the three phases of the system in input and their outputs are connected together and to the load. Only one Triac is operating at a time; hence, the load will be connected to one phase while the other two Triacs are off.

It is to be noted that no auxiliary commutation circuit is utilized in the STS. Let us assume, that Triac-1 was on and the load was connected to Phase-A. Once the load transfer command is received by the end-user controller to transfer the load from Phase-A to B, it will block the gate signals for Triac-1. However, Triac-1 will still continue to supply the load until its forward current falls below its holding current. Then, it will turn-off and gate signals can be applied to Triac-2. This lack of timing control for a Triac is the main drawback of the STS. Gate Turn-Off Thyristors (GTOs) could be applied instead. However, they have complex drivers, are more expensive and have higher conduction losses. These are the main reasons why Triacs were chosen for the STS.

E. *Reliability Issues*

In general, there are two main reliability concerns in the smart load transfer system. The first one is the failure probability (or lost data packets) in the master controller or ZigBee-based communication system. The second one is the failure probability of STS (including controller, semiconductors, PCB, etc.).

Failure in the master controller and the communication system will stop the smart load transfer system; however it will not have any excessively harmful effects. The network will become unbalanced over time but the consumers will remain connected.

Failure in STS can be discussed separately for failure in open-circuit and short-circuit conditions. If a STS fails in short circuit, the house no longer will be able to participate in load transfer scheme. The affected household remains connected. It is possible that the remaining STS units will still effectively maintain the network

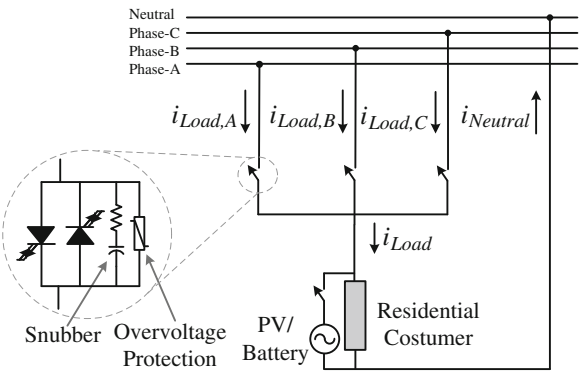


Fig. 2 Schematic diagram of a static transfer switch for residential loads

balance. If a STS fails in open-circuit, then the electricity supply to the house will be interrupted and the STS needs to be replaced immediately.

A survey on reliability issues on communication systems in smart grids and power electronics devices are presented in [16, 17].

F. *Load Transition Dynamic Issues*

Transferring a load from one phase to another results in dynamic input voltage variation for that load. ITI (formerly CBEMA) curve or some recently developed curves define the acceptable limits of the voltage that electrical appliances can work safely and is considered by the manufacturers in building and testing their products [18, 19]. If the voltage is kept within the acceptable region in ITI curve, the electrical appliances within residential premises should operate correctly and not suffer any damage.

The smart load transfer system operates in 15-min time intervals to minimise the nuisance effects of frequent load transfers. In addition, a deadband controller is also developed which activates the load transfer scheme if voltage unbalance is above an acceptable (desired) limit. Applying the deadband controller limits the load transfers only to the periods with unacceptable voltage unbalance in the network and therefore, the number of load transfers within the network is highly reduced. The discrete operation of load transfer as well as the deadband controller minimise the load transfer numbers and their peripheral effects on network voltage variations.

3 Load Transfer Decision Making Process

As discussed before, the master controller is responsible for selecting the candidate load(s) to be transferred from one phase to another to reduce the voltage unbalance along the feeder. In this section, the voltage unbalance calculation methods are discussed first. Some conventional voltage unbalance reduction methods are discussed and their limitations for the smart load transfer system are then highlighted. Later, a practical decision making process is discussed to be used in the smart load

transfer system for voltage unbalance reduction. Finally, a genetic algorithm-based optimization method is discussed for optimal selection of the candidate load(s) within the feeder.

3.1 Voltage Unbalance Definition

Voltage unbalance in the three-phase electric system is a condition in which the three phase voltages (V_A, V_B and V_C) differ in amplitude and/or do not have 120 degree phase difference. Several methods for calculation and interpretation of voltage unbalance are investigated in [20, 21].

Based on IEEE Std. 1996 [21] or IEC Std. 61000-4-27, voltage unbalance is defined from line voltages (V_{AB}, V_{BC} and V_{CA}) as

$$\% \; VU_{TD} = \left|\frac{V_-}{V_+}\right| \times 100 = \left|\frac{V_{AB} + a^2 V_{BC} + a V_{CA}}{V_{AB} + a V_{BC} + a^2 V_{CA}}\right| \times 100 \tag{1}$$

where $a = 1\angle 120°$ and V_- and V_+ are the negative and positive sequence of the voltage, respectively.

Majority of the researches nowadays utilize (1), referred to as Voltage Unbalance True Definition (VU_{TD}). Note that VU_{TD} can also be calculated from phase voltages. However, in [21], it was stated that (1) neglects the zero sequence component. In four-wire systems utilising a star connection of the loads and the transformer low voltage, the zero sequence components often makes a larger contribution to voltage unbalance than the negative sequence components. Hence, voltage unbalance in low voltage four-wire feeders is defined as [21]:

$$\% \; VU_{4-wire} = \frac{\sqrt{V_-^2 + V_0^2}}{|V_+|} \times 100 \tag{2}$$

This is referred to as percentage voltage unbalance in this chapter. It can be shown that for a four-wire three-phase low voltage feeder with equally sized phase and neutral conductors, $VU_{4\text{-}wire}$ is higher than VU_{TD}.

According to [22], the allowable limit for voltage unbalance is limited to 2 % in low voltage networks. Engineering Recommendation P29 in UK not only limits network voltage unbalance to 2 %, but also limits the voltage unbalance to 1.3 % at the load point [23].

3.2 Mathematical Methods for Voltage Unbalance Minimization

Two different mathematical methods for minimizing the voltage unbalance in one bus are discussed below.

A. *Sensitivity Analysis*

The symmetrical components for phase voltages are given as

$$\begin{bmatrix} V_+ \\ V_- \\ V_0 \end{bmatrix} = \frac{1}{3}\begin{bmatrix} 1 & a & a^2 \\ 1 & a^2 & a \\ 1 & 1 & 1 \end{bmatrix}\begin{bmatrix} V_A \\ V_B \\ V_C \end{bmatrix} \tag{3}$$

Replacing (3) in (2), voltage unbalance in a four-wire system can be simplified as

$$\%\ VU_{4-wire} = \sqrt{\frac{2\left(V_A^2+V_B^2+V_C^2\right)+2V_A.V_B\cos(\theta_{AB}+60)+2V_B.V_C\cos(\theta_{BC}+60)+2V_C.V_A\cos(\theta_{CA}+60)}{\left(V_A^2+V_B^2+V_C^2\right)+2V_A.V_B\cos(\theta_{AB}-120)+2V_B.V_C\cos(\theta_{BC}-120)+2V_C.V_A\cos(\theta_{CA}-120)}} \tag{4}$$

where $\theta_{AB} = \theta_A - \theta_B$, $\theta_{BC} = \theta_B - \theta_C$ and $\theta_{CA} = \theta_C - \theta_A$.

Voltage unbalance sensitivity versus dependent β can be calculated as [24]

$$S_\beta^{\%\ VU_{4-wire}} = \frac{\beta}{\%\ \mathrm{VU}_{4-wire}} \cdot \frac{\partial}{\partial\beta}(\%\ VU_{4-wire}) \tag{5}$$

where β is any of V_A, V_B, V_C, θ_A, θ_B and θ_C parameters. When calculating voltage unbalance sensitivity versus any of them, the other dependent parameters need to be constant. Finally, it can be simplified as

$$S_\beta^{VU\%} = \beta \times \frac{\frac{\partial}{\partial\beta}\left(V_-^2\right)+\frac{\partial}{\partial\beta}\left(V_0^2\right)-\frac{\partial}{\partial\beta}\left(V_+^2\right)\times\ \%VU_{4-wire}^2}{2\times\ \%VU_{4-wire}^2\times V_+^2} = \frac{1}{2}\left(S_\beta^{V_-^2+V_0^2}-S_\beta^{V_+^2}\right) \tag{6}$$

In this case, voltage unbalance sensitivity analysis will determine which β parameter has the most effect on voltage unbalance reduction.

The selected β parameter does not reveal the candidate load(s) to be transferred in a real low voltage feeder, with several different loads distributed in random locations and number in each phase. In addition, a load transfer in any bus will result in the voltage magnitude and angle change in all phases and all buses. Therefore, the assumption of independency between all the three phase voltage magnitudes and angles is invalid. Hence, voltage unbalance sensitivity analysis is not helpful.

B. *Vector Analysis*

Let us consider a three-phase balanced four-wire source with one single-phase to neutral load. The load is a constant current of I at power factor of $\cos\varphi$. First, let us assume the load is connected to Phase-A. Therefore, $i_A = I\angle-\varphi$ and the sequence components of current are

$$I_+ = I_- = I_0 = \frac{I}{3}\angle-\varphi.$$

Similarly, let us assume the load is now connected to Phase-B. The current in this phase is $i_B = I\angle -120^\circ - \varphi$ and the current sequence components can be written as

$$I_+ = \frac{I}{3}\angle -\varphi \qquad I_- = \frac{I}{3}\angle +120^\circ - \varphi \qquad I_0 = \frac{I}{3}\angle -120^\circ - \varphi\,.$$

In a similar way, when the load is connected to Phase-C, the current in this phase is $i_C = I\angle +120^\circ - \varphi$ and the sequence components are

$$I_+ = \frac{I}{3}\angle -\varphi \qquad I_- = \frac{I}{3}\angle -120^\circ - \varphi \qquad I_0 = \frac{I}{3}\angle +120^\circ - \varphi\,.$$

Now, for studying the effects of a load transfer from one phase to another phase, let us assume the load in Phase-A is now transferred to Phase-B. In this situation, the change in the current sequence components is

$$\begin{aligned}
\Delta I_+^{A\to B} &= -\frac{I}{3}\angle -\varphi + \frac{I}{3}\angle -\varphi = 0\\
\Delta I_-^{A\to B} &= -\frac{I}{3}\angle -\varphi + \frac{I}{3}\angle +120^\circ - \varphi = \frac{I}{\sqrt{3}}\angle +150^\circ - \varphi\\
\Delta I_0^{A\to B} &= -\frac{I}{3}\angle -\varphi + \frac{I}{3}\angle -120^\circ - \varphi = \frac{I}{\sqrt{3}}\angle -150^\circ - \varphi
\end{aligned} \tag{7}$$

In a similar way, let us assume the load connected to Phase-B is transferred to Phase-C. In this case, the change of current sequence components will be

$$\begin{aligned}
\Delta I_+^{B\to C} &= -\frac{I}{3}\angle -\varphi + \frac{I}{3}\angle -\varphi = 0\\
\Delta I_-^{B\to C} &= -\frac{I}{3}\angle +120 - \varphi + \frac{I}{3}\angle -120^\circ - \varphi = \frac{I}{\sqrt{3}}\angle -90^\circ - \varphi\\
\Delta I_0^{B\to C} &= -\frac{I}{3}\angle -120 - \varphi + \frac{I}{3}\angle +120^\circ - \varphi = \frac{I}{\sqrt{3}}\angle +90^\circ - \varphi
\end{aligned} \tag{8}$$

Similarly, the current sequence change when the load is transferred from Phase-C to A will be

$$\begin{aligned}
\Delta I_+^{C\to A} &= -\frac{I}{3}\angle -\varphi + \frac{I}{3}\angle -\varphi = 0\\
\Delta I_-^{C\to A} &= -\frac{I}{3}\angle -120^\circ - \varphi + \frac{I}{3}\angle -\varphi = \frac{I}{\sqrt{3}}\angle +30^\circ - \varphi\\
\Delta I_0^{C\to A} &= -\frac{I}{3}\angle +120^\circ - \varphi + \frac{I}{3}\angle -\varphi = \frac{I}{\sqrt{3}}\angle -30^\circ - \varphi
\end{aligned} \tag{9}$$

Equations (7)–(9) can be simplified as

$$\begin{bmatrix} \Delta I_+ \\ \Delta I_- \\ \Delta I_0 \end{bmatrix}_{A\to B} = \frac{I\angle -\varphi}{\sqrt{3}} \begin{bmatrix} 0 \\ 1\angle +150^\circ \\ 1\angle -150^\circ \end{bmatrix}$$
$$\begin{bmatrix} \Delta I_+ \\ \Delta I_- \\ \Delta I_0 \end{bmatrix}_{B\to C} = \frac{I\angle -\varphi}{\sqrt{3}} \begin{bmatrix} 0 \\ 1\angle -90^\circ \\ 1\angle +90^\circ \end{bmatrix} \tag{10}$$
$$\begin{bmatrix} \Delta I_+ \\ \Delta I_- \\ \Delta I_0 \end{bmatrix}_{C\to A} = \frac{I\angle -\varphi}{\sqrt{3}} \begin{bmatrix} 0 \\ 1\angle +30^\circ \\ 1\angle -30^\circ \end{bmatrix}$$

From (10), it can be seen that, when the load is transferred from one phase to another, the change in the positive sequence of current is zero. However, the negative and zero sequences change.

In a similar way, it can be proved that the load transfer from Phase-B to A will result in 180 degree phase transfer compared to load transfer from Phase-A to B for both of the negative and zero sequence changes. It is also similar for transfers from Phase-C to B and from Phase-A to C. The negative and zero sequence change for a load transfer between any two phases is shown in Fig. 3.

Example-1: Let us consider a four-wire system connecting a three-phase balanced voltage source to load terminals through equivalent impedance of Z_P for each phase. The neutral conductor is assumed to have an impedance of Z_N. Therefore, the sequence components of the impedance are

$$Z_+ = Z_- = Z_P \qquad Z_0 = Z_P + 3Z_N.$$

First, let us assume the system has three similar loads. Each load draws a constant current I with unity power factor (i.e. $\cos\varphi = 1$). Two of the loads are connected to Phase-A, one to Phase-B and C has no loads. Therefore, the load terminal voltages are

$$\begin{aligned} V_A &= V\angle 0 - 2Z_P I\angle 0 \\ V_B &= V\angle -120^\circ - Z_P I\angle -120^\circ \\ V_C &= V\angle +120^\circ \\ V_N &= (2I\angle 0 + I\angle -120^\circ)Z_N. \end{aligned}$$

Now, the phase voltages can be written as

$$\begin{aligned} V_{AN} &= V_A - V_N = V\angle 0 - 2Z_P I\angle 0 - (2I\angle 0 + I\angle -120^\circ)Z_N \\ V_{BN} &= V_B - V_N = V\angle -120^\circ - Z_P I\angle -120^\circ - (2I\angle 0 + I\angle -120^\circ)Z_N \\ V_{CN} &= V_C - V_N = V\angle +120^\circ - (2I\angle 0 + I\angle -120^\circ)Z_N \end{aligned}$$

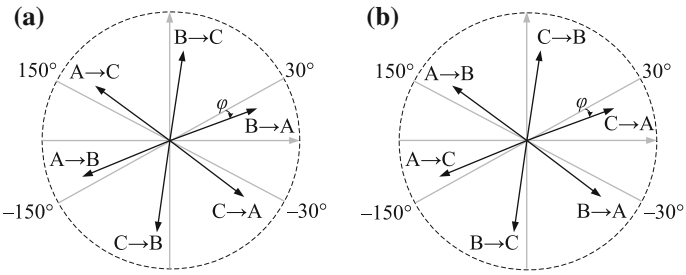


Fig. 3 **a** Change in current zero sequence for load transfer between phases. **b** Change in current negative sequence for load transfer between phases

which can be simplified as

$$\begin{bmatrix} V_{AN} \\ V_{BN} \\ V_{CN} \end{bmatrix} = \begin{bmatrix} 1 & 2Z_P + (a^2+2)Z_N \\ a^2 & a^2 Z_P + (a^2+2)Z_N \\ a & (a^2+2)Z_N \end{bmatrix} \begin{bmatrix} V \\ -I \end{bmatrix}.$$

The negative sequence voltage at load terminal is

$$V_- = 0.88\angle -19.1^\circ.I.Z_P.$$

For eliminating the voltage negative sequence, it is required to inject negative sequence current ($I_{-,inj}$) so that the negative sequence voltage becomes zero. The required negative sequence current for this purpose is

$$I_{-,inj} = -(0.88\angle -19.1^\circ.I).$$

For this case, based on the angle of the calculated $I_{-,inj}$, it can be stated that it is equal to a load transfer from Phase-A to C, as seen from Fig. 3b. For the same example, the zero sequence of voltage at load terminal is

$$V_0 = 0.57\angle -30^\circ.I.Z_0.$$

Now, similarly, the required amount of injected zero sequence current ($I_{0,inj}$) to eliminate V_0 must be

$$I_{0,inj} = -(0.57\angle -30^\circ.I).$$

For this case, based on the angle of the calculated $I_{0,inj}$, it can be stated that it is equal to a load transfer from Phase-A to C, as seen from Fig. 3a. Therefore, both of the diagrams of Fig. 2, imply that a load transfer from Phase-A to C will minimize voltage unbalance at load terminals. ΔΔΔ

Example-2: Now, let us assume there are three types of loads. All of them draw a constant current of I but with different power factors. Three loads with power factor of 0.98 ($\varphi = 10^\circ$) are connected Phase-A. One load with power factor of

0.93 ($\varphi = 20°$) is connected to Phase-B. Two loads with power factor of 0.86 ($\varphi = 30°$) are connected to Phase-C.

Now, the voltages at the load terminal are

$$V_A = V\angle 0 - 3Z_P I\angle -10°$$
$$V_B = V\angle -120° - Z_P I\angle -140°$$
$$V_C = V\angle +110° - 2Z_P I\angle +90°$$
$$V_N = (3I\angle -10° + I\angle -140° + 2I\angle +90°)Z_N.$$

In this case, it can be shown that the required injected zero and negative sequence currents are

$$I_{0,inj} = 0.78\angle -159°.I$$
$$I_{-,inj} = 0.39\angle -152°.I.$$

For this case, based on the angle of the calculated $I_{0,inj}$ from Fig. 3a, the equivalent action is load transfer is Phase-A to B. However, based on the angle of $I_{-,inj}$ from Fig. 3b, the equivalent action is the load transfer from Phase-A to C. In this example, one portion of the loads on Phase-A are to be transferred to Phase-B and another portion to Phase-C in order to compensate the negative and zero sequences simultaneously. ΔΔΔ

Example-2 demonstrated decision making on which load(s) to be transferred from one phase to another is not straight forward through vector analysis. On the other hand, Eq. (10) is based on the assumption of a balanced voltage source. However, in reality, it is quite possible that the voltages at any bus are already unbalanced due to other loads connected to the feeder before that bus.

3.3 Highly-Loaded to Low-Loaded with Deadband Control

Voltage unbalance at each bus is substantially proportional to the difference between the voltage magnitudes of three phases in that bus. Therefore, to reduce voltage unbalance in each bus, a method that equalises the amplitude of all three phase voltages is implemented. This can be achieved by transferring the load from the highly-loaded phase to the low-loaded phase at that bus. This process can be continued until the best voltage unbalance for all buses along the feeder is achieved. The variations in voltage unbalance should be monitored to prevent voltage unbalance increase due to an inappropriate load transfer. An exhaustive method is used for applying a load change from highly-loaded phase to low-loaded phase in each bus followed by calculation of voltage unbalance for all buses. The load transfer which has resulted in the best voltage unbalance for all buses in the network is chosen as the desired load transfer result. The flowchart of the control algorithm is shown in Fig. 4.

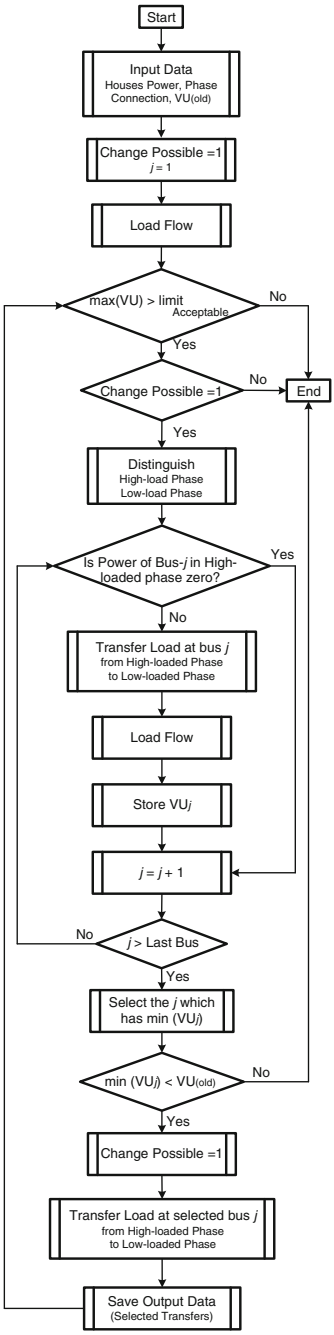


Fig. 4 Flow chart of the decision making process in smart load transfer system based on highly-loaded to low-loaded phase with deadband control

Although the result from this method is not globally optimum, it has the great advantage of few load transfers in each action of the smart load transfer system. It is to be noted that, in this method, voltage unbalance reduction along the feeder is the main objective and power mismatch among the three phases as well as voltage regulation along the feeder are improved consequently.

3.4 Genetic Algorithm-Based Optimization

The voltage unbalance can be reduced even further if a general optimization is carried for the system. For this purpose, a genetic algorithm-based optimization method is chosen [25]. The genetic algorithm optimization, utilized in this analysis, is discussed briefly below.

A. *Initialization*
For the genetic algorithm in this study, a population size of 100 is randomly selected. Each population (chromosome) is a vector with column number (gen) equal to house number. A non-binary gen is utilized in this analysis. For that, the value of each vector array (cell) can be 1, 2 or 3 which respectively represent Phase-A, B or C connection for each house. Each cell must have a value of 1, 2 or 3 at every time to represent the house is definitely connected to one phase.

B. *Selection*
A pair of individuals from the population is selected as a parent. In this analysis, the individuals with less voltage unbalance impact in the network have a higher probability to be selected as parents. Parents are utilized for defining the child generations in the next two sections.

C. *Crossover*
Crossover is an important operation and is used for exchanging information among the selected parent individuals from the last section. In this analysis, a one-point crossover with a randomly selected truncation point and a constant probability of 30 % is utilized.

D. *Mutation*
As another important operation, mutation is carried out over the selected parent individuals. In this analysis, a one-point mutation with a randomly selected truncation point and a constant probability of 5 % is used.

E. *Objective Function*
The objective function in this analysis is assumed to be as

$$OF_p = \beta_1.\max(VU_p) + \beta_2.\max\left(\Delta P_p\right) \tag{11}$$

where OF_p is the objective function calculated for each population p, VU_p is the vector of voltage unbalance along the feeder for each population p, ΔP_p is the

vector of three-phase power mismatch at the secondary side of the distribution transformer for each population p and β_1 and β_2 are respectively the penalty factors for voltage unbalance and power mismatch. The genetic algorithm parameters and the penalty factors need to be tuned to result in fast convergence and better figures for voltage unbalance and power mismatch among the three phases in the network under consideration.

F. *Stopping Rule*

In each iteration, the population resulting in the best objective function (i.e. minimum of voltage unbalance and power mismatch) is saved. The iterative procedure of generating new trials is terminated when no significant improvement is observed in the solutions (objective function values for new trials).

The flowchart of the genetic algorithm-based smart load transfer control system is shown in Fig. 5.

The main advantage of this method is finding the most suitable phase connection for the houses in each interval so that the objective function is minimized. However, this will result in a high number of load transfers in the system each time.

4 Analysis Methodology

A MATLAB-based program is developed to analyse the results of the smart load transfer system. The simulation comprised the main program, smart load transfer program and the unbalanced load flow program.

The low voltage residential distribution feeder of Fig. 1 is considered in this study. The distribution transformer is assumed to be three-phase, 100 kVA, 11/0.4 kV, 50 Hz, DYn type, with per-unit impedance of 4 %. The low voltage feeder is assumed to be an aerial three-phase four-wire radial system which is composed of four similar Mercury conductors (i.e. 7/4.5 AAC with $Z = 0.315 + j\ 0.259\ \Omega$/km [26]). The overhead line has a length of 400 m, distributed over 120 cm cross-arms, with ABCN vertical configuration [27]. It is assumed that the feeder supplies 30 houses through 10 buses, with equal separations of 40 m (i.e. 1 house per phase per bus). The utilized residential load profile in this analysis is based on the real residential data retrieved from the smart meters installed in a suburban area in Perth, Australia [28]. The load profile for each house for 1 day, in the considered network, is shown in Fig. 6. The loads are assumed to be as constant PQ loads within this study. From this figure, it can be seen that some loads have negative load demand at some periods which is due to their PV power generation.

In addition, for studying the effectiveness of the discussed smart load transfer system, dynamic simulations are carried out using PSCAD/EMTDC.

An unbalanced sweep forward–backward load flow [29] is considered and integrated into the developed model. The load flow calculates the voltages along

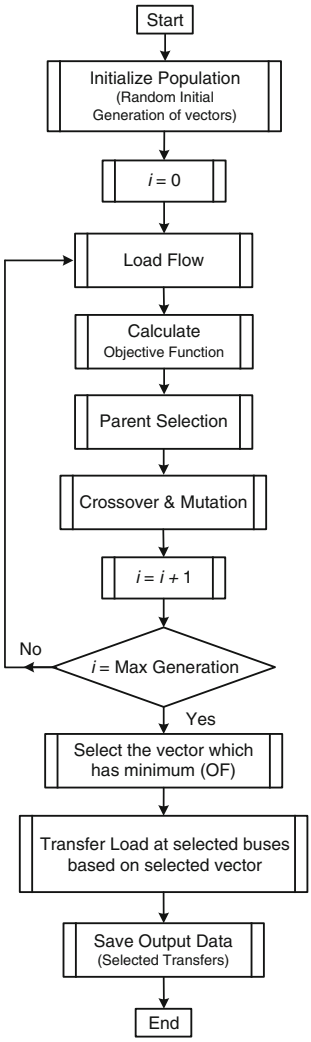


Fig. 5 Flow chart of the decision making process in smart load transfer system using genetic algorithm-based optimization

the feeder which are used later for calculating voltage unbalance along the feeder. The load flow algorithm is described in Chap. 10.

5 Simulation Results

Using the developed MATLAB-based program, the performance of the smart load transfer system for the considered network in Fig. 1 is demonstrated below:

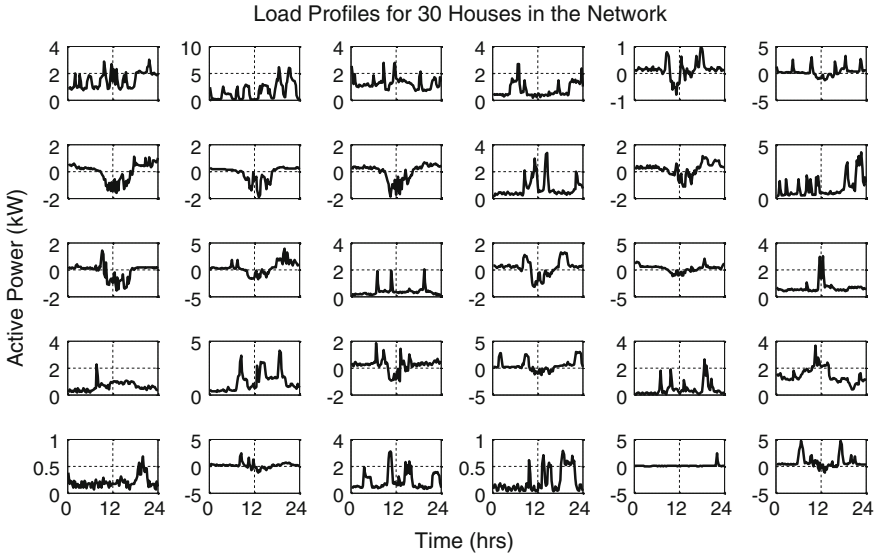


Fig. 6 Load profile for the 30 houses participating in smart load transfer system

A. *Highly-loaded to Low-loaded Transfer without Deadband Control*
First, let us assume the deadband controller is deactivated and the load transfer scheme is running for any voltage unbalance in the network. In such a case, the three-phase voltage profile of the feeder before and after three load transfers is shown in Fig. 7a. The voltage unbalance profile along the feeder is shown for the initial case and after three load transfers in Fig. 7b. Table 1 shows the numerical results for voltage unbalance in each bus in addition to the location of load transfers.

For the network under consideration, Fig. 8a shows the maximum of voltage unbalance along the feeder in each 15-min time interval before and after smart load transfer system application. This figure shows that the smart load transfer system is highly successful in reducing the voltage unbalance along the feeder in the 24-h period. For the studied data, the maximum of network voltage unbalance was 2.23 % which was reduced down 0.16 % after load transfer was applied in that period. After the smart load transfer system is applied, the maximum of experienced voltage unbalance along the feeder, in 24-h period, is 0.77 %.

The minimum voltage along the feeder for each time interval before and after application of the smart load transfer system is shown in Fig. 8b. This figure shows that the minimum voltage of the feeder is improved by applying the smart load transfer system. As an example, the minimum voltage in the feeder in the case without smart load transfer system was 0.97 pu which increased to 0.99 pu.

In a similar way, the maximum voltage along the feeder for each time interval before and after the application of the smart load transfer system is shown in Fig. 8c. This figure shows that the maximum voltage of the feeder is reduced after

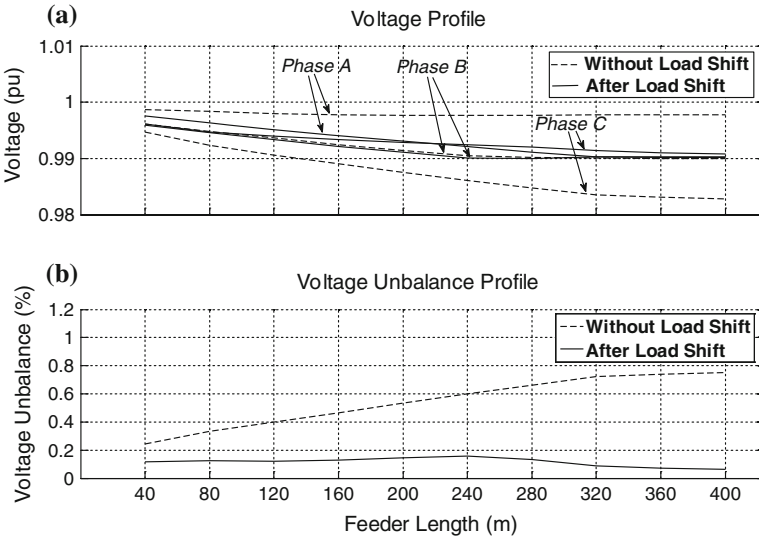


Fig. 7 **a** Network voltage profile before and after smart load transfer. **b** Network voltage unbalance profile before and after smart load transfer

Table 1 Voltage unbalance (%) in each bus along the feeder before and after each load transfer in highly-loaded to low-loaded transfer control algorithm

Bus number	1	2	3	4	5	6	7	8	9	10
Without LT	0.24	0.33	0.39	0.46	0.53	0.60	0.66	0.72	0.74	0.75
After first load transfer at bus-8 Phase C → A	0.09	0.11	0.10	0.10	0.11	0.12	0.16	0.22	0.23	0.24
After second load transfer at bus-7 Phase C → B	0.07	0.06	0.03	0.03	0.06	0.09	0.13	0.17	0.17	0.17
After third load transfer at bus-6 Phase A → B	0.10	0.09	0.09	0.10	0.12	0.14	0.14	0.12	0.10	0.09

utilising the smart load transfer system. It is to be noted that these changes are the consequence of voltage unbalance reduction and are not controlled directly.

It is highly desirable to achieve better results with fewer load transfers in each switching case. Figure 8d shows the total number of load transfers in each switching case. It can be seen that total number of load transfers was between one and nine in each switching case. This means that in the worst case, the load transfer was applied to maximum of 30 % of the houses.

It is interesting to investigate if there were any houses in the network which had more load transfers applied to them. Figure 8e shows the total number of load transfers for each house at each bus of the network. From this figure, it can be seen that all houses participated in the load transfer scheme but some had more transfers than others. In Fig. 8f, the phase connection of each house is shown during the 24-h period. In this figure, Phase-A, B and C are respectively labelled as Phase-1, 2 and 3.

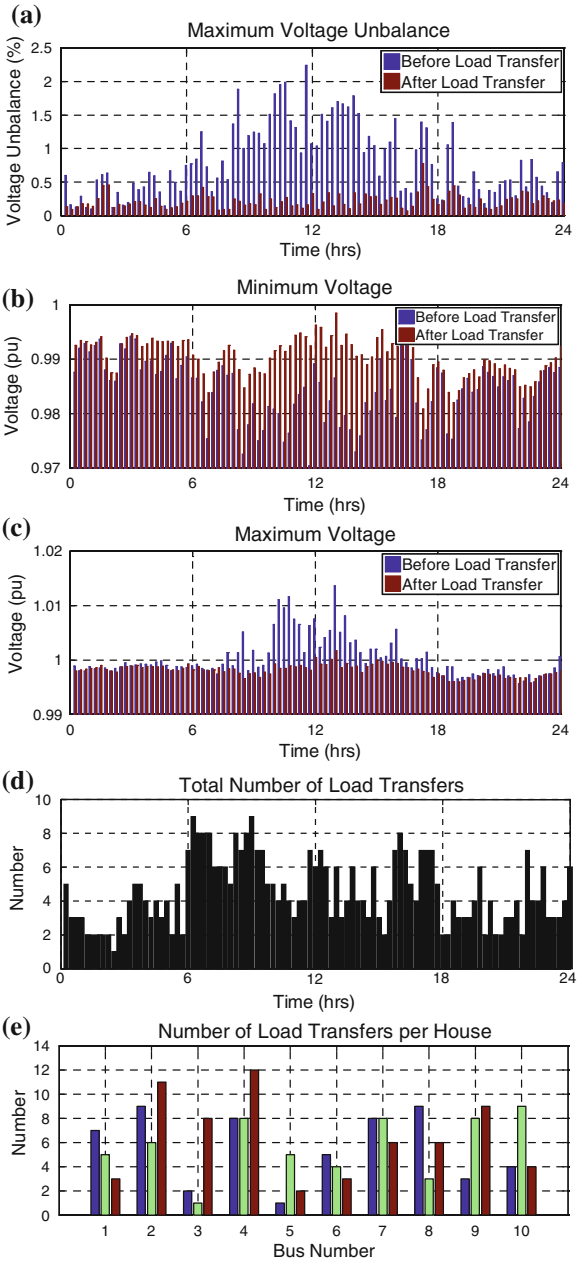
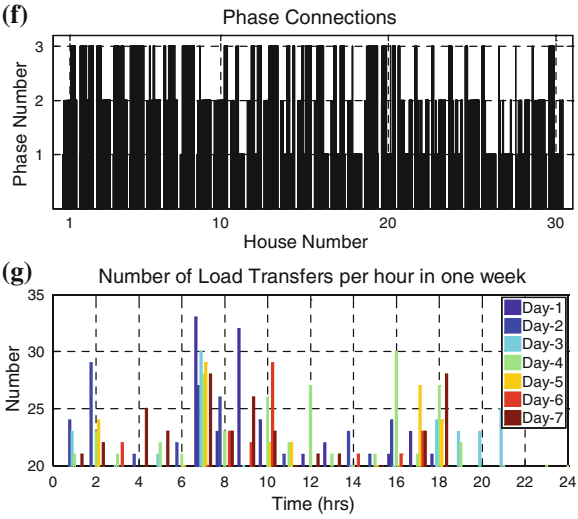


Fig. 8 Results of smart load transfer system application in 24-h: **a** Maximum voltage unbalance in the feeder before and after load transfer. **b** Minimum voltage magnitude in the feeder before and after load transfer. **c** Maximum voltage magnitude in the feeder before and after load transfer. **d** Total number of load transfers in each switching case (15-min time intervals). **e** Total number of load transfers per each house. **f** Demonstration of phase connection of each house in the network. **g** Total number of load transfers in the network per hour in 7 days

Fig. 8 (continued)



It is also interesting to analyse if there was any specific time per day that the number of load controls were higher. For this, the above study has been extended for a week using the available 7-day real load profile data. Figure 8g shows the number of total load transfers in the network in each hour for a 7-day period. Only the load transfer numbers higher than 20 per hour are demonstrated in this figure. This figure shows that the number of load transfers is almost higher at 7.00–10.00 am every day. This is the time of the first peak in the network load demand per day. However, not many load transfers are observed at 16.00–18.00 pm which was time of major peak in network load demand. The total number of load transfers per hour and per switching mode is given for the 7-day period separately in Table 2. From the statistics in this table, it can be seen that at every switching case, a maximum of 30 % of the houses (i.e. 10 houses) and a minimum of 3 % (i.e. one house) had participated in the load transfer scheme on average. From the statistics, it can also be seen that the maximum number of load transfers in an hour is about three times of the minimum number of load transfers in an hour every day. In addition, on average, a minimum number of two and a maximum number of 13 load transfers was observed per house in a 7-day period.

B. *Load Transfer with Deadband Control*

As discussed in Sect. 4, for reducing the number of load transfers in a network, a deadband controller was designed which activates the load transfer scheme only once the voltage unbalance in the network is higher than its setting (i.e. the acceptable/desired level of voltage unbalance in the network). Now, let us assume in the network of Fig. 1, the deadband controller is active and has a setting of 0.5, 1 and 1.5 % for voltage unbalance. The results of applying the highly-loaded to low-loaded load transfer algorithm are shown in Fig. 9 for each setting of the deadband controller separately. As seen from this figure, as the setting for the

Table 2 Load transfer statistical data in a 7-day period

Day		1	2	3	4	5	6	7
Number of load transfers per each switching case	Maximum number	9	9	10	10	9	9	9
	Minimum number	1	2	1	2	2	1	2
Number of load transfers per each hour	Maximum number	33	29	30	20	29	29	28
	Minimum number	8	11	13	13	10	12	13
Number of load transfers per house	Maximum number	12	12	12	15	12	14	13
	Minimum number	1	3	2	3	1	2	2

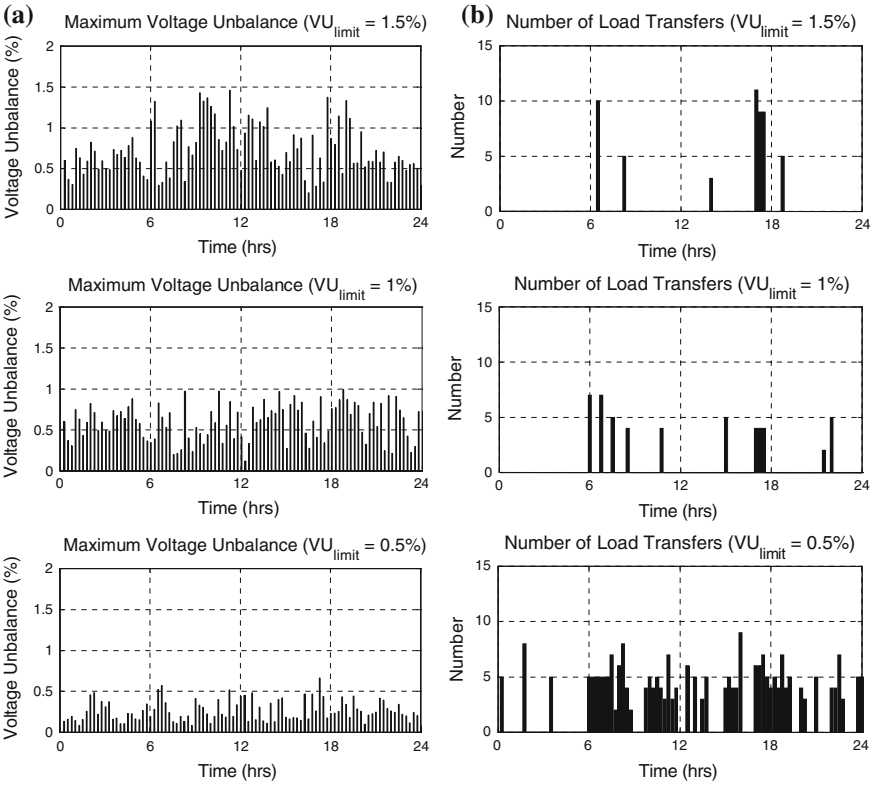


Fig. 9 Comparison of 0.5, 1 and 1.5 % settings for deadband control for: **a** Maximum voltage unbalance in feeder after load transfer. **b** Total number of load transfers in each switching case

acceptable voltage unbalance in the network is increased from 0.5 to 1.5 %, the number of load transfers is decreased significantly subject to higher voltage unbalance presence in the network. As an example, for limiting the acceptable voltage unbalance in the network from 1.5 to 0.5 % (i.e. one-third), the required total number of load transfers in the network in a 24-h period was increased from 52 to 254 transfers (i.e. five times more).

C. *Genetic Algorithm-Based Optimization*

Another study is carried out to demonstrate the performance of genetic algorithm-based smart load transfer system with the objective function considering voltage unbalance and power mismatch in three phases, as described in Sect. 4. The results of this study for the network of Fig. 1 are shown in Fig. 10.

The results are given for the network after applying the smart load transfer system scheme based on two different optimizations. In the first optimization, only voltage unbalance is minimized (i.e. $\beta_1 = 1$ and $\beta_2 = 0$); while in the second one, voltage unbalance and power mismatch among the three phases are minimized together (i.e. $\beta_1 = 1$ and $\beta_2 = 20$). It is to be noted that the above penalty factors were tuned to result in better results of *OF* for a crossover probability of 30 % and mutation probability of 5 % in the network under consideration. Figure 10 shows that voltage unbalance reduction using genetic algorithm optimization is more than the reduction by highly-loaded to low-loaded algorithm. However, this will lead to a higher number (i.e. 60 % more) of load transfers in the network. In addition, this figure shows that the power mismatch among the three phases is reduced significantly in genetic algorithm-based load transfer scheme, when this parameter is considered in the objective function as in (11).

D. *Dynamic Simulation Results*

As discussed before, the load transfer should not result in unacceptable input voltages for the residential customers. Input voltage variations should be within the acceptable region of the ITI curve. For studying the dynamic performance of the smart load transfer system using STS, the system in Fig. 2 is modeled in detail in PSCAD/EMTDC. It is assumed that a single-phase 2 kW load with power factor of 0.95 is supplied by a three-phase 240 V RMS voltage through a STS.

First, let us assume the load is being supplied from Phase-A. At $t = 0.5$ s, a command is received from the master controller to the end-user controller to transfer the load to Phase-B followed by another command at $t = 1$ s to transfer the load to Phase-C. The load instantaneous voltage and current waveforms are shown in Fig. 11a, b while their RMS values are shown in Fig. 11c, d. The instantaneous current waveform is scaled up in this figure for better presentation.

Once the control command is received at $t = 0.5$ s, the controller will block the Phase-A Triac and will de-block Phase-B Triac. However, the load current is not zero; therefore, the conducting thyristor will still continue passing the current until it falls below its holding current. This happens at $t = 0.51$ s. However, it can only start to conduct the current at $t = 0.57$ s, once the voltage of Phase-B is in its positive half cycle and its relevant thyristor is activated. This will cause a slight drop in voltage RMS which is within acceptable range in the ITI curve.

At $t = 1$ s, another load transfer command is received to transfer the load to Phase-C. At this time, the current is in its negative half cycle while Phase-C voltage is in its positive half cycle. Therefore, Phase-B Triac has to still conduct the current until the current falls below its holding current. This happens at

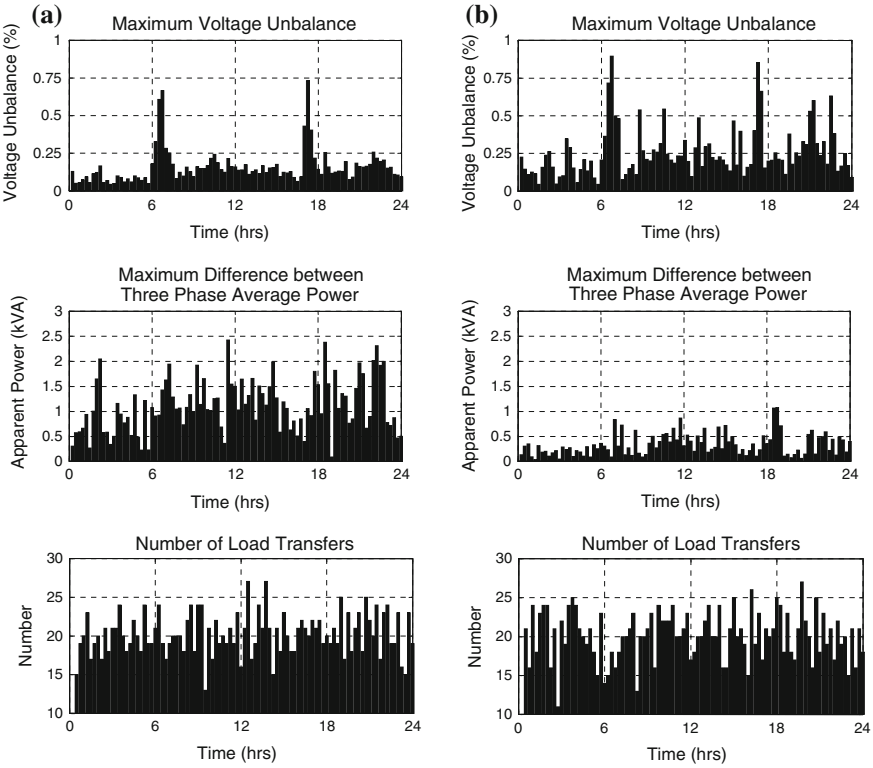


Fig. 10 Comparison of results of maximum voltage unbalance, maximum power difference from average power among three phases and number of load transfers after load transfer scheme based on Genenetic Algorithm for: **a** Voltage unbalance minimization. **b** Voltage unbalance and power mismatch minimization

$t = 1.05$ s, therefore, the conducting Triac is turned off and since the voltage of Phase-C is in its negative half cycle and its relevant thyristor is activated, then it starts to conduct the current. This results in a slight increase in voltage RMS which is again within the acceptable range of the ITI curve.

As seen from Fig. 11c, the voltage variation during a load transfer transition period is below 10 % and does not last more than 100 ms; which falls within the safe region of the ITI curve.

Now, let us assume the load is 1 kW while a PV generating 2 kW is connected within the residential promises (i.e. a negative 1 kW load demand). The PV is modeled based on the voltage–current characteristic of PV cells with a single-phase inverter as discussed in [30]. Figure 11e shows the load active power demand while a similar load transfer command is applied. The simulation results verify the successful dynamic performance of the STS-based smart load transfer system for residential applications.

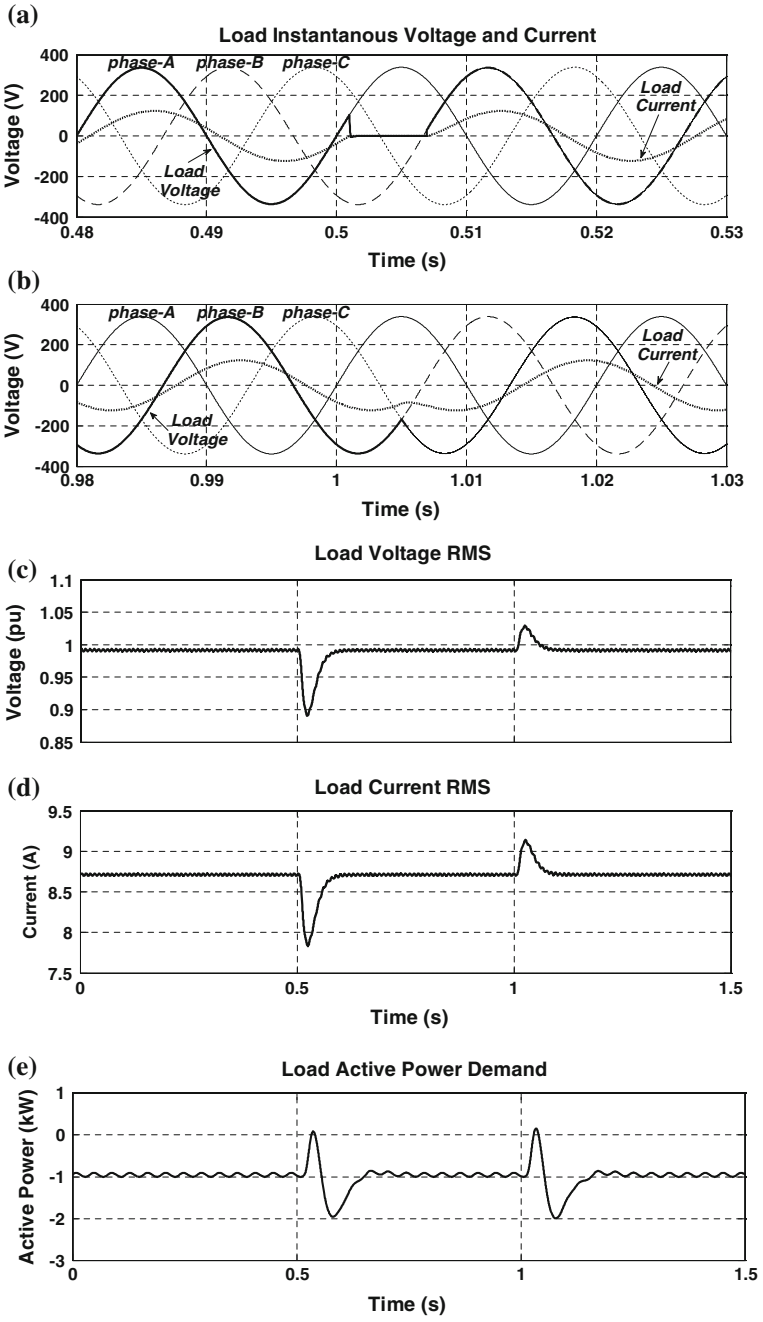


Fig. 11 Dynamic simulation results: **a** Load instantanouse voltage for a load transfer at $t = 0.5$ s from Phase-A to B. **b** Load instantanouse voltage for a load transfer at $t = 1$ s from Phase-B to C. **c** Load voltage RMS. **d** Load current RMS. **e** Load active power demand in presence of a PV

6 Conclusion

A smart dynamic residential load transfer system was discussed in this chapter. The supply of each house could be transferred from one phase to another based on the commands from the master controller using a static transfer switch. The master controller defines the candidate load(s) to be transferred based on the discussed highly-loaded to low-loaded phase transfer algorithm, while keeping the number of load transfers to minimum. The smart load transfer system operated in 15-min time intervals and a deadband controller was developed to reduce the number of load transfers, by allowing an acceptable level of voltage unbalance within the network. A genetic algorithm-based optimization can also be utilized for voltage unbalance and power mismatch reduction through a larger number of load transfers. The effectiveness of the smart load transfer system was demonstrated using a MATLAB-based simulation for a typical Australian low voltage distribution feeder using the load profile data available from smart meters. The dynamic analysis of the static transfer switch demonstrated that the voltage variation during each load transfer falls within the acceptable levels of the input voltage for the electrical appliances, based on the ITI curve.

References

1. A. Ghosh, G. Ledwich, *Power Quality Enhancement using Custom Power Devices* (Kluwer Academic Publishers, Boston, 2002)
2. T.A. Short, *Electric Power Distribution Handbook* (CRC Press, Boca Raton, 2004)
3. F. Shahnia, R. Majumder, A. Ghosh, G. Ledwich, F. Zare, Voltage imbalance analysis in residential low voltage distribution networks with rooftop PVs. Electr. Power Syst. Res. **81**(9), 1805–1814 (2011)
4. F. Shahnia, A. Ghosh, G. Ledwich, F. Zare, Voltage unbalance sensitivity analysis of plug-in electric vehicles in distribution networks, in *21st Australasian Universities Power Engineering Conference* (*AUPEC*) (2011), pp. 1–6
5. F. Shahnia, A. Ghosh, G. Ledwich, F. Zare, Voltage correction in low voltage distribution networks with rooftop PVs using custom power devices, in *37th Annual IEEE Industrial Electronics Society Conference* (*IECON*) (2011), pp. 991–996
6. R.J. Sarfi, M.M.A. Salama, A.Y. Chikhani, A survey of the state of the art in distribution system reconfiguration for system loss reduction. Electr. Power Syst. Res. **31**(1), 61–70 (1994)
7. C. Lueken, P.M.S. Carvalho, J. Apt, Distribution grid reconfiguration reduces power losses and helps integrate renewables. Energy Policy **48**, 260–273 (2012)
8. S. Jazebi, B. Vahidi, Reconfiguration of distribution networks to mitigate utilities power quality disturbances. Electr. Power Syst. Res. **91**, 9–17 (2012)
9. B. Amanulla, S. Chakrabarti, S.N. Singh, Reconfiguration of power distribution systems considering reliability and power loss. IEEE Trans. Power Delivery **27**(2), 918–926 (2012)
10. T.H. Chen, J.T. Cherng, Optimal phase arrangement of distribution transformers connected to a primary feeder for system unbalance improvement and loss reduction using a genetic algorithm. IEEE Trans. Power Syst. **15**(3), 994–1000 (2000)

11. A. Ghosh, Performance study of two different compensating devices in a custom power park. IEE Proc. Gener. Transm. Distrib. **152**(4), 521–528 (2005)
12. H. Pezeshki, P. Wolfs, Correlation based method for phase identification in a three phase LV distribution network, in *22nd Australasian Universities Power Engineering Conference* (*AUPEC*) (2012), pp. 1–7
13. S.S.S.R. Depuru, W. Lingfeng, V. Devabhaktuni, N. Gudi, Smart meters for power grid: challenges, issues, advantages and status. Renew. Sustain. Energy Rev. **15**(6), 2736–2742 (2011)
14. J. Gao, Y. Xiao, J. Liu, W. Liang, C.L.P. Chen, A survey of communication/networking in Smart Grids. Future Gener. Comput. Syst. **28**(2), 391–404 (2012)
15. M.E. Kantarci, H.T. Mouftah, Supply and load management for the smart distribution grid using wireless networks, *in IEEE Japan–Egypt Conference on Electronics, Communications and Computers* (*JEC–ECC*) (2012), pp. 145–150
16. D. Niyato, W. Ping, E. Hossain, Reliability analysis and redundancy design of smart grid wireless communications system for demand side management. IEEE Wireless Commun. **19**(3), 38–46 (2012)
17. Y. Shaoyong, A. Bryant, P. Mawby, X. Dawei, L. Ran, P. Tavner, An industry-based survey of reliability in power electronic converters. IEEE Trans. Ind. Appl. **47**(3), 1441–1451 (2011)
18. G.T. Heydt, R. Ayyanar, R. Thallam, Power acceptability. IEEE Power Eng. Rev. **21**(9), 12–15 (2001)
19. S. Elphick, V. Smith, The 230 V CBEMA curve-preliminary studies, in *20th Australasian Universities Power Engineering Conference* (*AUPEC*) (2010), pp. 1–6
20. A.K. Singh, G.K. Singh, R. Mitra, Some observations on definitions of voltage unbalance, in *39th North American Power Symposium* (*NAPS*) (2007), pp. 473–479
21. M.T. Bina, A. Kashefi, Three-phase unbalance of distribution systems: Complementary analysis and experimental case study. Int. J. Electr. Power Energy Syst. **33**(4), 817–826 (2011)
22. IEEE Recommended Practice for Monitoring Electric Power Quality, IEEE Standard 1159–1995
23. Planning Limits for Voltage Unbalance in the United Kingdom, The Electricity Council, Engineering recommendation P29, 1990
24. D.C. Garcia, A.L.F. Filho, M.A.G. Oliveira, O.A. Fernandes, F.A. do Nascimentoa, Voltage unbalance numerical evaluation and minimization. Electr. Power Syst. Res. **79**, 1441–1445 (2009)
25. J. Zhu, *Optimization of Power System Operation* (Wiley, New York, 2009)
26. Olex Aerial Catalogue, http://www.olex.com.au/Australasia/2012/OLC12641_AerialCat.pdf
27. Western Power Distribution Construction Standards Handbook, (2007), http://www.westernpower.com.au/documents/networkcontractors/distributionconstruc/we_n5115114_v2_distribution_construction_standards_handbook_.pdf
28. Y. Li, P. Wolfs, Hybrid model for residential loads in a distribution system with high PV penetration. IEEE Trans. Power Syst. **28**(3), 3372–3379 (2013)
29. W.H. Kersting, *Distribution System Modeling and Analysis* (CRC Press, Boca Raton, 2012)
30. F. Shahnia, R. Majumder, A. Ghosh, G. Ledwich, F. Zare, Operation and control of a hybrid microgrid containing unbalanced and nonlinear loads. Electr. Power Syst. Res. **80**(8), 954–965 (2010)

Part V
Industrial Applications

The Optimization of Computational Stock Market Model Based Complex Adaptive Cyber Physical Logistics System: A Computational Intelligence Perspective

Bo Xing

Abstract This chapter makes an attempt to address three critical issues that, from a computational intelligence perspective, will arise when computational stock market model (CSMM) based complex adaptive cyber physical logistics system (CACPLS) is implemented in the future supply network. The chapter starts with an introduction and background description about the necessity of introducing the CSMM-based CACPLS; then the focal problems (i.e., developing investment strategy, predicting stock price, and controlling extreme events) of this chapter is stated in the problem statement section; a detailed description about our approaches, i.e., training artificial neural network via particle swarm optimization, genetic algorithm for stock price forecasting, and agent-based modeling and simulation for preventing extreme events, together with three example studies can be found in the subsequent proposed methodology sections; right after this, the potential research directions regarding the key problems considered in this chapter are highlighted in the future trends section; finally, the conclusions drawn at the last section closes this chapter.

Keywords Complex adaptive system · Artificial stock market model · Cyber physical logistics system · Artificial neural network · Genetic algorithm · Multi-agent system

1 Introduction

The efficiency of logistics system (LS) is a key component for survival of companies in today's hyper-competitive business environment. However, upgrading LS's management in the context of international supply chain is not a smooth

B. Xing (✉)
Department of Mechanical and Aeronautical Engineering, Faculty of Engineering, Built Environment and Information Technology, University of Pretoria, Pretoria, South Africa
e-mail: bxing2009@gmail.com

Z. H. Khan et al. (eds.), *Computational Intelligence for Decision Support in Cyber-Physical Systems*, Studies in Computational Intelligence 540,
DOI: 10.1007/978-981-4585-36-1_12,

process. Several authors (e.g., [111], Wu [104]) have argued that uncertainties in terms of supply, production and demand are the important feature on the contemporary business environment and thus make LS much more vulnerable. Others (e.g., [24, 95]) pointed out that LS can be, from a complexity science-based perspective [45, 46], regarded as complex adaptive system (CAS). As a result, LS is unmanageable, particularly when the phenomena of extreme events occurred. In this chapter, we proposed the concept of complex adaptive cyber physical logistics system (CACPLS) in response to these challenges confronted by companies.

Briefly, the remainder of this chapter is organized as follows. Section 2 includes background and literature review about CAS, cyber physical system, and CACPLS, which is followed by a description of our focal problems in Sect. 3. Then, the proposed computational intelligence (CI) methodology and the experimental study for each particular research question is detailed in Sects. 4, 5, and 6, respectively. Next, the future research directions are highlighted in Sect. 7. Finally, the conclusion is drawn in Sect. 8.

2 Background of Complex Adaptive Cyber Physical Logistics System

2.1 What is Complex Adaptive System?

Before we apply CAS theory, we must first understand it. Complexity theory is well established in a range of disciplines, such as economics [65], ecosystems [57], health care [50, 98], signaling networks [44], mathematical optimization [101], and in some areas of logistic and supply chain [24, 66, 77, 78, 95, 105], to name but a few. Adaptation has been identified as a fundamental characteristic of CAS [42] that has been appearing under the term adaptive management in the resource management literature. Also, systems theory is an analytical framework that has been widely used across scientific disciplines since World War II [21]. It has been providing a variety of conceptual and mathematical methods to gain insight into the behavior of a set of component that interact through the flow of matter and energy and through feedback mechanisms [39].

The term CAS organized as a system with seven basic elements that emerges over time into a coherent form, adapting and organizing itself without any singular entity controlling or managing it [43]. The first four concepts are aggregation, nonlinearity, flow and diversity, which represent certain characters of agents, are very important in the adaptation and evolution process, while the other three concepts are tagging, internal models and building blocks, which represent the mechanisms of agents for communication with the environment [70]. A similar viewpoint has been presented by [24], who listed the attributes in three broad groups as environment (dynamism, and rugged landscape), internal-mechanisms (deals with agents, self-organization and emergence, connectivity and dimensionality), and

co-evolution (quasi equilibrium and state changes, non-linear changes, and non-random future). Overall, CAS is extremely flexible allowing for any form of dynamics and evolution to be considered. The difficulty comes in defining structures capable enough to exhibit the features of interest, while still remaining amenable to mathematical analysis [34, 115]. A further problem is that this incomplete description can lead to results being proven for one type of structure representation not necessarily holding for others [36]. Therefore, if sense is to be made of the structure and function of CAS, a fully integrated approach will be necessary.

2.2 What is Cyber Physical System?

Cyber physical system (CPS) is a promising new class of engineered systems that presents the integrations of computation with physical environments. Just as the internet help humans to interact with one another, CPS help us to interact and control with the physical world. So far, there are a wide range of technologies that have been involved in building the CPS. One of these technologies, radio frequency identification (RFID), is able to provide an automatic identifying and tracking link between multiple objects at the same time, without the need for individual scanning [88], has been thoroughly studied. Examples includes a range of large-scale engineered systems such as manufacturing [73], closed-loop supply chain (CLSC) [116], and inventory systems [67]. Furthermore, CPS is characterized by decentralization and autonomous behavior of their elements. In addition, such systems evolve through collective adaptation and reconfiguration of their structures.

2.3 Why Complex Adaptive Cyber Physical Logistics System (CACPLS)?

Being in transient phase of rapid growth, the LS is quite complicated and deeply affected by the novel information and communication technologies (ICT). As a result, it is inevitably falling into the category of CPS in which all objects are equipped with sensors (e.g., RFID) in order to record the environment of the object such as where they are, which other things are in the vicinity, and what happened to them in the past (See Fig. 1 for illustration).

For example, Ivanov and Sokolov [48] identified an inter-disciplinary perspective and modeling tools in the context of collaborative CPS. Furthermore, in order to achieve an effective control on LS, the researchers of [24, 95, 96] pointed out that CAS and multi-agent system (MAS) have been extensively applied to LS domain so far. Clearly, those technological advantages bring many benefits to the

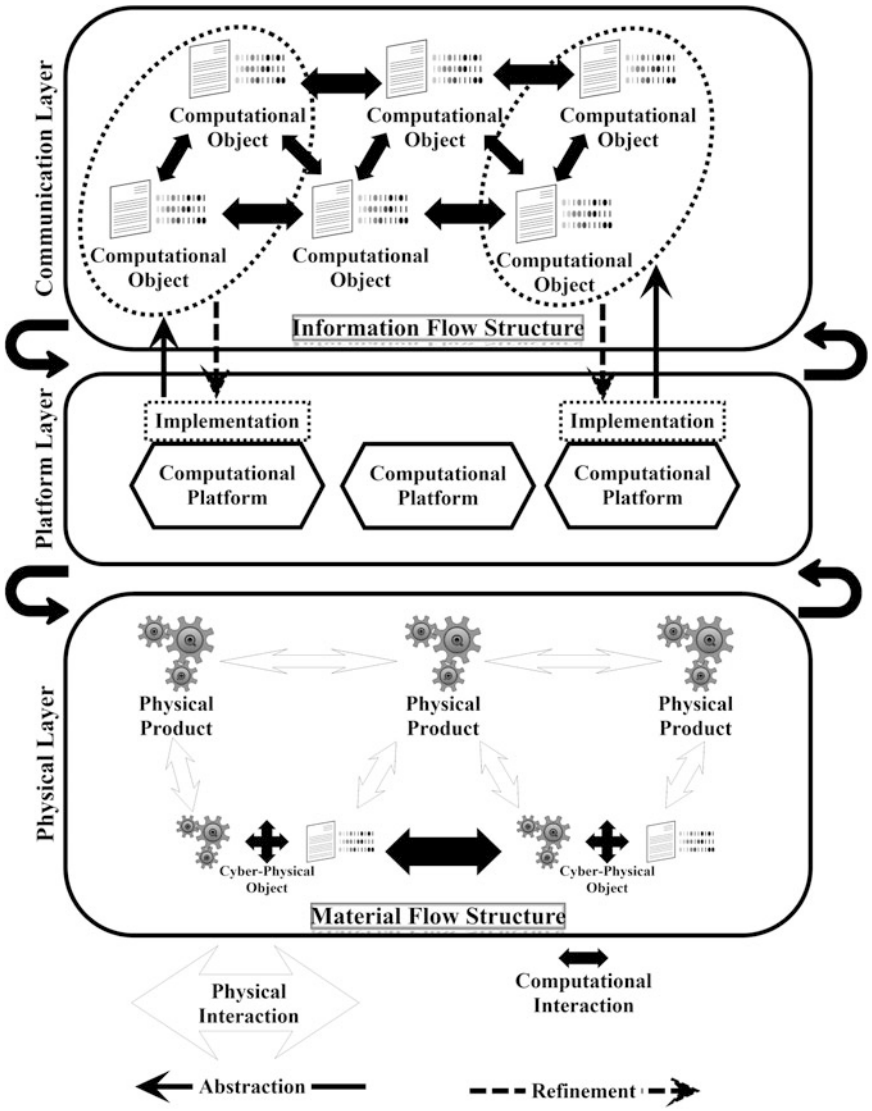


Fig. 1 CACPLS diagram

LS such as they make the LS safer and more efficient, performing tasks more accuracy with reduced cost, and providing real-time information of operational transactions to corporate-level information systems for the purpose of better planning and execution. As a result, in this chapter we combined those advance ideas and proposed a novel concept called complex adaptive cyber physical logistics system (CACPLS) to address planning and operation challenges of LS.

Such systems will depend on a cyber-infrastructure of sensing, communication, and control devices, which means that the LS is entirely possible to adapt themselves automatically in terms of supply, production and demand.

2.4 Computational Stock Market Model Based CACPLS

Computational stock markets [55, 75] (or artificial stock markets) are market simulations populated with artificially intelligent electronic agents that fill the roles of traders. These agents can use heuristics, rules, and various CI approaches to make trading decisions. That means, the market allows a fine-grained level of experimental control that is not available in real markets. Thus, data obtained from artificial market experiments can be compared to the predictions of theoretical models and to data from real-world markets, and the level of control allows one to examine precisely which settings and conditions lead to the deviations from theoretical predictions usually seen in the behavior of real markets [26]. A summary of computer-simulated markets with individual adaptive agents to explain observed market phenomena can be found in [53].

3 Problem Statement

One of the major challenges for logistics practitioners is to develop a network structure and collaboration mechanism that can facilitate adaptive, flexible and synchronized behaviors in a dynamic environment. However, researchers are still in the early stages of investigating the general principles that govern the birth, growth and evolution of supply networks with complex network structure and mechanisms for collaboration. In other words, if logistics operations are regarded as being complex, it is inappropriate to consider models developed under paradigms based on beliefs of linearity, stability, homogeneity, and perfect rationality as producing the best possible explanations and understanding of turbulent contexts. In this chapter, we argue that logistics networks should be treated as a CAS since they are characterized by self-organizing properties causing emergent system-wide effects. With this recognition, we will adopt computational stock market model (CSMM) [66] based CACPLS. This model can, in turn, be used to improve the tactical and operational decision-making of logistics operations. Although adopting CSMM sounds like a promising way to design CACPLS, it is not a panacea for solving all logistics related problems. Some issues outlined below need to be paid an extra attention before reaching a CSMM-based CACPLS.

3.1 Research Question 1 (RQ1): Developing Better Investment Strategy

First, in many financial decision problems, such as portfolio optimization or hedging, an investor wants to find an optimal investment strategy to achieve a certain goal, for example, to maximize a utility function or to minimize risk. The development of new successful investment strategies, or the improvement of methodologies to product new successful investment strategies can be a profit business venture [28].

But, how are investment strategies developed? The answer can vary across asset classes. In the case of stocks and corporate bonds, traditional fundamental analysis entails analyzing the corporation, the quality of the assets, and the specifics of the securities issued. Such analysis is usually carried out through the study of traditional quantitative indicators emphasizing value (various price/earnings metrics), financial stability (liquidity ratios), and qualitative opinions such as management depth and expertise and market dominance. The goal of such analysis is to determine what is the real, intrinsic value of a security, and then to compare that value to the price being offered in the market. When a discrepancy between the intrinsic value and the market price exists, there is a chance to profit by buying securities believed to be undervalued and selling securities believed to be overvalued. On the other hand, investment strategies can also be based on qualitative factors such investing in "green" companies (e.g., remanufacturing firms [107] with a superior focus on corporate social relations and alternative production methodologies [28].

Often investment strategies are based on both qualitative and quantitative factors, specializing in a specific market niche and then analyzing specific securities based on various managerial variables. Therefore new investment strategies are usually developed by a combination of innovative hypothesizing and empirical research. Generally a human uses various financial analysis tools to discover a repeated discrepancy between intrinsic value and market price, and then formulates an investment strategy to take advantage of this perceived discrepancy. The search for new investment strategies is often carried out by thousands of finance professionals around the world, and has the potential to yield huge profits if found. For this reason, it is always worth thinking not only about individual potential investment strategies, but also about refining the process through which new strategies are developed.

Nevertheless, there are two primary obstacles existing in the process of developing new investment strategies by human: firstly, human thought processes must choose what variables are significant and worth spending time to analyze. This process can thus be biased both by traditional investing philosophy, and by the lack of human conceptualization of potential relationships among different variables; secondly, there exists a bottleneck in human's capability to process and analyze large data sets. An analyst might be interested in potentially investing in thousands of publicly listed companies, but would never have time to thoroughly analyze all of

their public statements [28]. The issue caused by such situation force us to come up the following research question: *RQ1—How investment strategy can be effectively developed for a better implementation of CSMM-based CACPLS?*

3.2 Research Question 2 (RQ2): Predicting Future Stock Price

Second, a stock price is a typical non-stationary stochastic process having no constant mean level over time for it to remain in equilibrium. Forecasting is an attempt to predict how a future event (such as stock return) will occur. The main objective of forecasting the occurrence of this event is for decision makers to make better decisions. In finance, forecasting is an important activity. Although according to efficient market hypothesis (EMH), classical economics tells us that the market cannot be predicted, forecasting of the financial markets, also known as financial forecasting, still has been attracting the interest of many market professionals, academics and even 'amateur' investors for many years.

Traditionally, there have been two forecasting methods that are widely used: the fundamental analysis and the technical analysis [49]. Fundamental analysis applies the tenets of economy foundation theory, which believes that various fundamental factors count. Technical analysis, on the other hand, insist that prices discount all information, and that studying the price trends can help trades forecast future price movement. Today, technical analysis has become widespread in the financial markets. While it's narrow form seeks to forecast the direction of price movements of target financial variables from their corresponding past priced and volume data. In this chapter, we make an attempt to address the following research question: *RQ2—How stock price can be reasonably predicted in the context of CSMM-based CACPLS?*

3.3 Research Question 3 (RQ3): Controlling Harmful Extreme Events

Third, a big swing in the prices of shares in the stock market is always a big story on the evening news. People often speculate on where the market is heading and get very excited when they can brag about their latest "big killing," but they become depressed when they suffer a big loss. The attention the market receives can probably be best explained by one simple fact: It is a place where people can get rich—or poor—quickly [69]. Therefore, rarely does a day just go by that the stock market isn't a major news item. Especially in recent years, we have witnessed huge swings in the stock market. The 1990s were an extraordinary decade for stocks: the Dow Jones and S&P 500 indexes increased more than 400 %, while

the tech-laden NASDAQ index rose more than 1,000 %. By early 2000, both indexes had reached record highs. Unfortunately, the good times did not last, and many investors lost their shirts. Starting in early 2000, the stock market began to decline: the NASDAQ crashed, falling by over 50 %, while the Dow Jones and S&P 500 indexes fell by 30 % through January 2003 [69]. These considerable fluctuations in stock prices affect the size of people's wealth and as a result may affect their willingness to spend. In [62], the author defined a financial extreme as the minimum daily return or the maximum daily return of a stock market index over a given period. Therefore it would be good if we can have a mechanism to prevent these extreme events and keep the markets in a well-organized condition. Among the most important players in financial markets throughout the world are central banks, the government authorities in charge of monetary policy. Central banks' actions affect interest rates, the amount of credit, and the money supply, all of which have direct impacts not only on financial markets, but also on aggregate output and inflation [69]. Typically three policy tools are often used in practice to manipulate the money supply, i.e., open market operations, changes in discount lending, changes in reserve requirements. Some brief introductions about these policies are provided as below [69]:

- Open market operations are the most important monetary policy tool, because they are the primary determinants of changes in interest rates and the monetary base, the main source of fluctuations in the money supply. Open market purchases expand reserves and the monetary base, thereby raising the money supply and lowering short-term interest rates. Open market sales shrink reserves and the monetary base, lowering the money supply and raising short-term interest rates.
- Discounting is a particularly effective way to provide reserves to the banking system during a banking crisis because reserves are immediately channeled to the banks that need them most. In addition to its use as a tool to influence reserves, the monetary base, and the money supply, discounting is important in preventing financial panics. Using the discount tool to avoid financial panics by performing the role of lender of last resort is an extremely important requirement of successful monetary policymaking.
- Changes in reserve requirements affect the money supply by causing the money supply multiplier to change. A rise in reserve requirements reduces the amount of deposits that can be supported by a given level of the monetary base and will lead to a contraction of the money supply. A rise in reserve requirements also increases the demand for reserves and raises the federal funds rate. Conversely, a decline in reserve requirements leads to an expansion of the money supply and a fall in the federal funds rate. The main advantage of using reserve requirements to control the money supply and interest rates is that they affect all banks equally and have a powerful effect on the money supply. The fact that changing reserve requirements is a powerful tool, however, is probably more of a curse than a blessing, because small changes in the money supply and interest rates are hard to engineer by varying reserve requirements. Another disadvantage of using reserve requirements to control the money supply and interest rates is that raising the

requirements can cause immediate liquidity problems for banks with low excess reserves. When the Fed has raised these requirements in the past, it has usually softened the blow by conducting open market purchases or by making the discount window more available, thus providing reserves to banks that needed them. Continually fluctuating reserve requirements would also create more uncertainty for banks and make their liquidity management more difficult.

Overall, no matter what kind of monetary policy is actually conducted, the ultimate goals are like keeping the stability of various financial markets. In the context of CSMM-based CACPLS, as we note in the earlier sections, smarts parts also face some possibility of extreme events in the form of too many parts buying space on a ship, then arriving at the dock to find the ship full and, thus are left stranded on the dock. Unsurprisingly, the probability of occurrence of extremes can have a large impact on economic development because intelligence products are then main objectives of intelligent transportation system for the businesses. This concern leads us to the third research question: *RQ3—How can extreme events are largely controlled when a logistics market becomes a CSMM-based CACPLS?*

4 Proposed Methodology for RQ1

Witnessed by these potential limitations of human in developing investment strategies, there is no doubt why LeBaron chose artificial neural network (ANN) as a new investment strategies developing mechanism in his CSMM [54]. The objective of ANN, as it pertains to trading rules, is to create a security trading decision support systems, which, ideally, is fully automated and triggered by both quantitative and qualitative factors. Furthermore, in a recent survey [92], the authors argued this choice (i.e., employing ANN in the development of new investment strategies) seems to be a right one and concluded that ANNs has the ability to extract useful information from large set of data and therefore plays an important role in stock market. Nevertheless, in practice, only through an effective training, ANN can become a promising solution in this regard. To address this issue, the particle swarm optimization research toolbox (PSORT) and the ANN training add-in for PSORT, written by George Evers [30, 31] and Tricia Rambharose [83], respectively, are proposed in this section for training ANN. The implementation of the experiment is conducted in a MATLAB environment.

4.1 PSO Research Toolbox

The PSORT is designed to handle continuous, single-objective optimization problems [31]. This simple toolbox allows researchers to solve new problems, or

improving the PSO. In order to add constraints, one simple approach for adding constraints to the particle swarm algorithm was presented. Particles are re-initialized until satisfying all constraints, and personal and global bests are updated only with feasible positions. This ensures that all starting points and all points of attraction are feasible without adding much complexity to an otherwise computationally simple algorithm. Alternatively, we can also enhance velocity updates and position updates, or vary the inertia weight non-linearly. For more details regarding how to apply PSORT to meet your own research needs, interested readers please refer to [31].

4.2 ANN Training Add-In for PSO Research Toolbox

The ANN training add-in for the original PSORT aims to allow an ANN to be trained using the PSO technique. This add-in works like a bridge or interface between MATLAB's NN toolbox and the PSORT. In this way, MATLAB's NN functions can call the NN add-in, which in turn calls the PSORT for NN training. The core principle behind this add-in is to treat each PSO particle as one possible solution of weight and bias combinations for the NN. The PSO particles therefore move about in the search space aiming to minimize the output of the NN performance function [83]. In terms of implementing this ANN training add-in, interested reader are referred to [83] for more instructions.

4.3 Experimental Study

The add-in works through the following functions created for ANN training with PSO [83]:

- **main.m**: This is a simple function used as an example of creating a neural network (NN) and using the PSO algorithm as the training function. Very basic input and target values are used, similar to the examples given in MATLAB's Neural Network toolbox help file. A two layer NN is created and NN parameters are set. The NN is first simulated and then trained. For comparative analysis of results, the NN settings and results are displayed by calling respective functions created for this purpose. Finally, a function is called to plot the NN performance results vs. epochs to enable more meaningful analysis of results.
- **trainpso.m**: This PSO training function aims to be the interface between the Matlab NN toolbox and the PSO Research Toolbox. For consistency, the formatting and structure of other NN training algorithms are followed. Training here is based on the idea that all weight and bias values are determined using a swarm optimization approach.
- **first_pso**: This variable is in the trainpso.m function as a flag to indicate to the PSO toolbox the first time the PSO algorithm is run. This is necessary to prevent

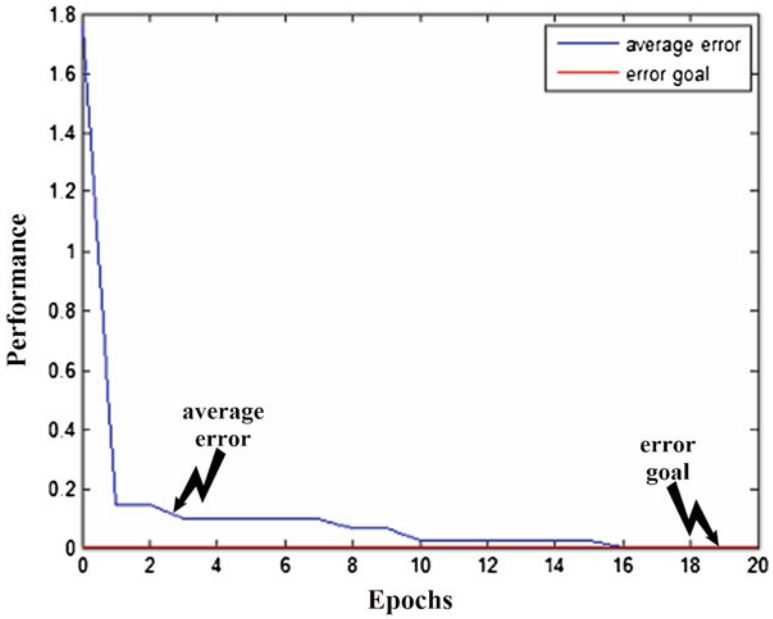


Fig. 2 Trainpso epochs versus performance

redundant PSO parameter validation and display that is needed only if the PSORT is not being used for ANN training. Initially, the value of this variable is set to "1" and after the first PSO trial its value is automatically set to "0" in RegPSO_main.m.

- **plot_epochs.m**: This is a script to plot NN epochs versus performance at the end of the ANN training session. ANN epoch number is given on the *x*-axis and the NN performance is given on the *y*-axis. For PSO training of an ANN, each PSO iteration corresponds to a NN epoch. For added analysis, the NN training goal is shown as a horizontal line. Figure 2 below gives a screenshot of the plot generated from plot_epochs.m.

5 Proposed Methodology for RQ2

Proponents of financial technical analysis have thus made serious attempts in the past decade to apply a variety of statistical models, and more recently, CI techniques to test the predictability of economic and stock market returns and volatility. Among the most popular and successful ones, we can find is ANN which is a very well-exploited CI technique in financial forecasting, while the other CI approaches used for financial forecasting such as genetic algorithm (GA) is less explored in the literature. Therefore, the aim of this section is to explore the applications of GA in financial forecasting.

5.1 Genetic Algorithm

Genetic algorithm (GA), introduced by Holland in 1975 [41], is a stochastic search technique. It has been successfully implemented in many combinatorial optimization problems. The concept of GA is based on the process of natural evolution. Typically GA starts with an initial set of possible solutions called population. Each individual in the population is called a chromosome, referred to as a solution to the problem at hand. Then some individuals are selected to be parents to produce offspring via a crossover operator. The evolution of new generations is carried out by breading the pairs of existing chromosomes in the current population using different genetic operators such as crossover, mutation, and inversion. Both offspring and parent chromosomes are then compared against the fitness function. The fitness function serves as the objective function and evaluates the quality of each chromosome. Only chromosomes with the highest fitness level will survive to form the next generation. Figure 3 illustrates a standard evolution process in GA.

5.2 Experimental Study

Considering the stock price forecasting problem, the model of the stock market that interacts with the investors is stochastic, dynamical non-linear and is not available [60]. The GA is programmed in MATLAB 7.6; and executed on Inter (R) Core(TM) i7-2040 M CPU @ 2.80 GHz computer with 2 GB RAM and Windows 7 Enterprise Operating System. The simulation results are illustrated in Fig. 4.

6 Proposed Methodology for RQ3

In order to control extreme events in stock market, on the contrary to conventional control policy tools which are currently regulating most stock markets, we resort to agent based modeling and simulation (ABMS) approach for stabilizing CALSs. ABMS represents a new paradigm in modeling and simulation of dynamic systems distributed in time and space. In particular, an agent performs given tasks automatically using inter-collaboration or negotiation with other agents on behalf of a human (such as buy and sell stock) on the basis of real-time connectivity. As the author of [53] put it, financial markets are one of the most important applications for agent based modeling because issues of price and information aggregation and dissemination tend to be sharper in financial settings, where objectives of agents are usually clearer.

Further, the availability of massive amounts of real financial data allows for comparison with the results of agent based simulations. For these reasons, agent technology is regarded as one of the best candidates for artificial stock market

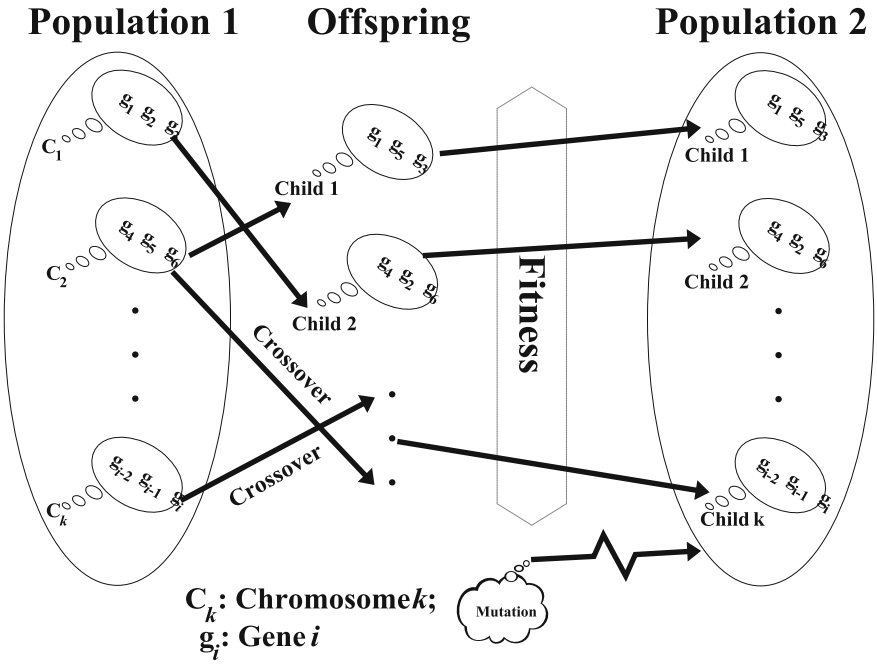


Fig. 3 A standard evolution process in GA

management. The basic idea in our proposed approach, which was inspired by [94], is to employ stochastic cellular automata model to generate dynamics of emergent stock market price as a result of the interaction between traders. After introducing socially integrated robot trading agents, the stock price dynamics could be controlled, so as to make the market more Gaussian.

6.1 Stochastic Cellular Automata Model

Cellular automata (CA) [72] are mathematical idealizations of physical systems in which space and time are discrete, and physical quantities take on a finite set of discrete values, and especially suitable for complex systems. A typical CA system comprises of four components namely: cells, states, neighborhood and rules. Cells are governed by transition rules that are iteratively applied, and generally only consider the states of the neighboring cells. The CA models can be applied in various areas to model the processes in physical, chemical, biological system, urban system [93], stock markets [102], and to study the self-organization and self-reproduction by the emergence of coherent interaction structures. Recently, to make the system wide or specific to certain situation, the CA models need to account for the external behaviors that also drive the changes which are not

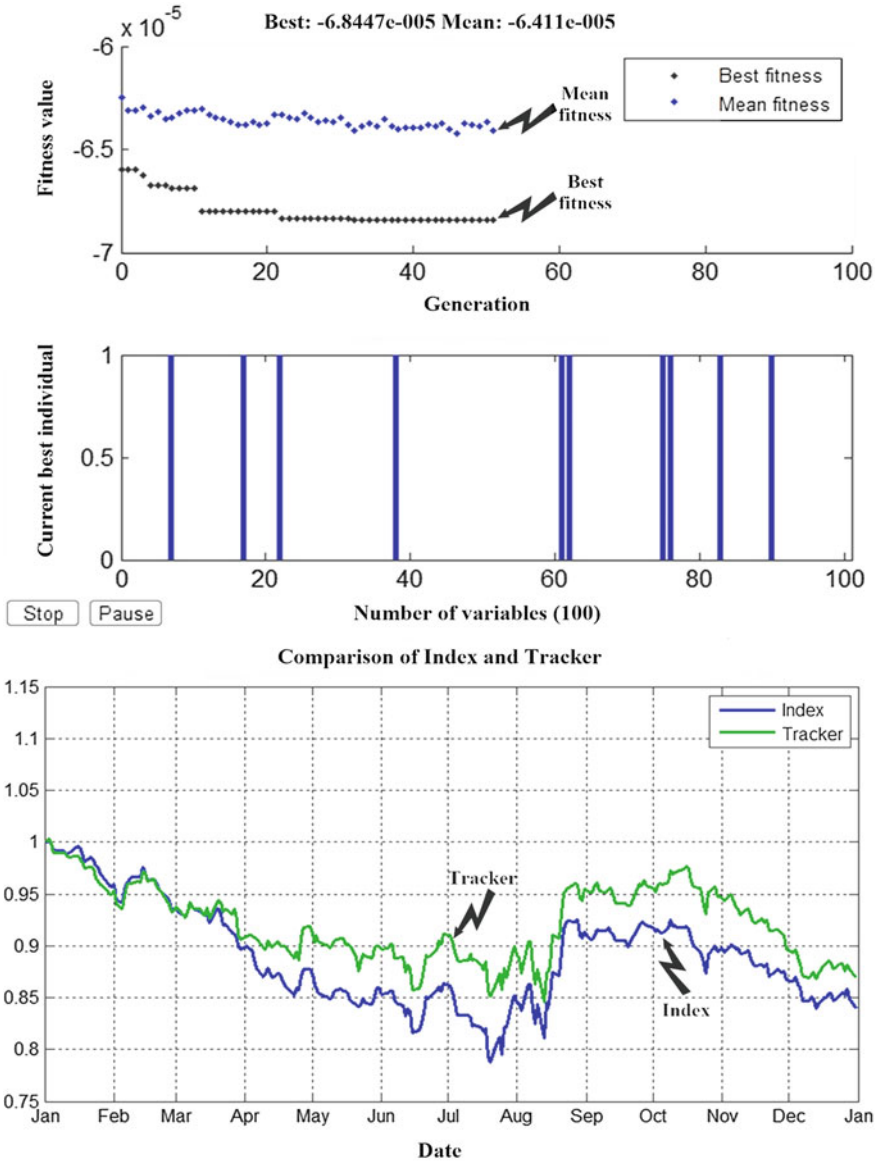


Fig. 4 Simulation results of predicting future stock price using GA

evolved in the transition rules of the automata. In the light of this statement, the new wave of research is driving towards integrating some new types with the CA models, such as agent CA [19] and ants CA [47]. The focus here is on to develop the framework for integrating the CA models and ABMs validating the same for the use case.

Initially, there are only human traders, and subsequently, robot trading agents enter the market. The traders are represented by cells on a two-dimensional $L \times L$ grid. There are N traders who can either buy or sell only one share, and these are mutually exclusive states. At any given time step t, the population of traders N is divided into two distinct groups of buyers $N_B(t)$ and sellers $N_S(t)$.

The probability of a trader to choose to buy at time t, $\pi^B(t)$, is in Eq. (1) [94]:

$$\pi^B(t) = S_1^B(t)^{\omega(t)} \cdot S_F^B(t)^{1-\omega(t)} \tag{1}$$

where $S_1^B(t) \in [0, 1]$ and $S_F^B(t) \in [0, 1]$, respectively, are probabilities of buying, based on imitation and on the fundamentals; and $w(t) \in [0, 1]$ is the weight ascribed on imitation. When $w(t) = 0$ the choice is based only on the fundamentals; when $w(t) = 1$ the choice is based only on imitation; and when $w(t) \in (0, 1)$ the choice mixes both strategies.

On the other hand, the probability of a trader choosing to sell, $\pi^S(t)$, is in Eq. (2) [94]:

$$\pi^S(t) = 1 - S_1^B(t)^{\omega(t)} \cdot S_F^B(t)^{1-\omega(t)} \tag{2}$$

In addition, we assume that the willingness to buy (sell) at t increases as a function of the number of neighbors who have already bought (sold) at the previous step $t-1$. Thus, the probability of imitation at time t, $S_1^B(t)$, is given by Eq. (3) [94]:

$$S_1^B(t) = \left[\frac{\left(N_B^H(t-1)\right)^k}{\left(N_B^H(t-1)\right)^k + \left(N_S^H(t-1)\right)^k} \right] \tag{3}$$

where, $N_B^H(t-1)$ and $N_S^H(t-1)$, respectively, are the number of buying neighbors and selling neighbors at $t-1$; and $k \in [1, \infty)$ is a parameter controlling the intensity of the response. When$k = 1$, the probability of buying is proportional to the number of neighbors who have previously bought; this characterizes a weak linear response. When $k > 1$ there is a quorum response, (i.e., the probability of exhibiting a particular behavior is an increasing function of the number of actors already performing the behavior), because the probability of buying increases once the quorum is met. Here, the quorum size is determined by the number of selling neighbors at the previous time step.

The probability of buying based on the fundamentals is then given by Eq. (4) [94]:

$$S_F^B(t) = \frac{e^{\lambda(\bar{F} - P(t-1))}}{e^{\lambda(\bar{F} - P(t-1))} + e^{-\lambda(\bar{F} - P(t-1))}} \tag{4}$$

where, the fundamental value $\bar{F}$ assumed as a positive constant; $P(t-1)$ is the stock price at $t-1$; and λ is a positive parameter modulating a trader's response, based on the price difference. If a trader perceives the stock price as being lower than the fundamental value $\bar{F} - P(t-1) > 0$, he tends to buy; and vice versa. When $\bar{F} - P(t-1) = 0$, the decisions of buying and selling are equally probable.

The weight ascribed to either strategy, $\omega(t)$ (that is, imitating or following the fundamentals), is endogenous, and depends on the size of the deviation of the current stock price from its fundamental value [94]:

$$\omega(t) = \frac{1}{1 + \mu[P(t-1) - \bar{F}]^2} \tag{5}$$

where, μ is a positive parameter, tracking the speed at which the strategy switches from imitation to fundamental. As, $[P(t-1) - \bar{F}]^2 \to \infty$, $\omega(t) \to 0$, and the fundamental strategy grows in importance; as $[P(t-1) - \bar{F}]^2 \to 0$, $\omega(t) \to 1$ and imitation is preferred.

We now introduce robot trading agents into the model. The robots are socially integrated into the group of human traders to control self-organized market returns. We assume that humans and robot agents are perceived as congeners and influence one another in the same way. Robot agent behavior is intentionally set to an anti-imitation rule to counteract the imitative human trader behavior as described by Eq. (3). The robot agents adopt such a contrarian behavior using the majority principle after considering the neighboring cells at the previous time step. Thus, the probability, $\pi_R^B(t)$, of a robot trading agent to choose to buy at time t is shown in Eq. (6) [94]:

$$\pi_R^B(t) = \left[\frac{N_S^H(t-1)}{N_B^H(t-1) + N_S^H(t-1)}\right] \tag{6}$$

where $N_B^H(t-1)$ and $N_S^H(t-1)$, respectively, are the number of buying neighbors and selling neighbors at $t-1$. According to Eq. (6), if $\pi_R^B \geq \frac{1}{2}$ a robot trader will choose to buy at time t; if $\pi_R^B < \frac{1}{2}$ the robot trader will choose to sell.

6.2 Agent-Based Modeling and Simulation

With the intense research in the realm of ABMS domain, scores of tools are developed for building ABMS. In this research, we choose one of them called NetLogo [103] because it was user-friendly and supported extensive documentation for building models. In the NetLogo parlance the agents are conceived as "turtles", the sense of state is through "patches" and the worldview through the "observer" [103]. Patches are similar to the notion of cell in CA with the regular lattice structure. Each cell in the CA terminology corresponds to the patch in the NetLogo parlance. Thus, the notion of space is based on regular lattice structures of square cells and agents are simulated to move over a cellular space.

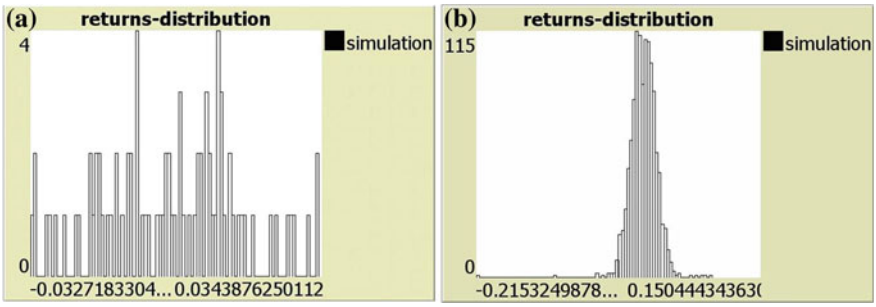


Fig. 5 Returns-distribution—(**b**) with and (**a**) without robot

6.3 Experimental Study

In order to run the experiments, we adopted the setup of the basic model: $\lambda = 1$, $k = 9$, and $\mu = 2 \times 10^{10}$. When there are no robot controllers are added, we can see through Fig. 5a that the returns is random distributed and environment is rather chaotic, while after we adding a certain number of robot controllers, the returns become more Gaussian distributed (see Fig. 5b).

7 Future Trends

7.1 Future Trends for RQ1

In this study, the standard PSO is introduced for training ANN. During the past few years, several hybridization versions of PSO have been proposed in the literature and enjoy a superior performance than original PSO. Among these variants, PSOGSA [68] (original PSO in combination with a newly developed gravitational search algorithm [18, 20, 22, 23, 27, 29, 32, 33, 37, 38, 40, 51, 52, 58, 59, 61, 74, 76, 84–87, 90, 91, 112, 117] seems to be an powerful alternative. As such, a possible future work in this regard would be testing the suitability of using PSOGSA for ANN training.

7.2 Future Trends for RQ2

In addition to employ the GA algorithm mentioned in this chapter, a potential research direction is to apply other innovative CI methods on the targeted question. The examples of innovative CI methods are such as superbug algorithm [14], bat algorithm [25, 56, 71, 110], blind, naked mole-rats (BNMR) algorithm [97], bean

optimization algorithm [99, 100, 113, 114], harmony search algorithm [4–8, 10–13, 15–17], artificial physics optimization (APO) [35, 106], river formation dynamics algorithm [79–82], particle collision algorithm [2, 3, 63, 64, 89], artificial chemical reaction [9, 109], and gases Brownian motion optimization [1]. Interested readers are referred to [108] for a more detailed description.

7.3 Future Trends for RQ3

Although the proposed ABMS method is efficient in controlling extreme events that might happen in CALS environment, it does not mean that the method is applicable to all situations. Meanwhile, the current parameter adjustment is done by manual. Therefore the future research can be focus on: First, finding out in what scenario the extreme event might be out of control and come up with a corresponding solution; Second, figuring out the best parameter choice for all scenarios (if it exists) or an alternative parameter selection strategy.

8 Conclusions

In this chapter, we first introduced the CACPLS concept into the context of logistics networks and supply chain management. Then, we identified three main research questions related to establishing a CSMM (i.e., computational stock market model) based CACPLS: developing better investment strategies, predicting future stock price, and controlling harmful extreme events. Three different approaches were proposed in this chapter to deal with our focal questions, namely, training ANN through PSO research toolbox, using GA to predict the stock price, and prevent extreme events via ABMS approach. The experimental studies conducted in this chapter demonstrated the suitability of our proposed methodologies.

References

1. M. Abdechiri, M.R. Meybodi, H. Bahrami, Gases Brownian motion optimization: an algorithm for optimization (GBMO). Appl. Soft Comput. **13**(5), 2932–2946 (2013). http://dx.doi.org/10.1016/j.asoc.2012.03.068
2. A. Abuhamdah, M. Ayob, Hybridization multi-neighbourhood particle collision algorithm and great deluge for solving course timetabling problems. Paper presented at the 2nd Conference On Data Mining and Optimization, (Selangor, 27–28 Oct 2009a), pp. 108–114
3. A. Abuhamdah, M. Ayob, *Multi-neighbourhood particle collision algorithm for solving course timetabling problems.* Paper presented at the 2nd Conference On Data Mining and Optimization (Selangor, 27–28 Oct 2009b), pp. 21–27

4. S. Afshari, B. Aminshahidy, M.R. Pishvaie, Application of an improved harmony search algorithm in well placement optimization using streamline simulation. J. Petrol. Sci. Eng. **78**, 664–678 (2011)
5. M.A. Al-Betar, I.A. Doush, A.T. Khader, M.A. Awadallah, Novel selection schemes for harmony search. Appl. Math. Comput. **218**, 6095–6117 (2012)
6. M.A. Al-Betar, A.T. Khader, A harmony search algorithm for university course timetabling. Ann. Oper. Res. **194**(1), 3–31 (2012)
7. M.A. Al-Betar, A.T. Khader, F. Nadi, *Selection mechanisms in memory consideration for examination timetabling with harmony search.* Paper presented at the Annual Conference on Genetic and Evolutionary Computation (GECCO) (Portland, 7–11 July 2010), pp. 1203–1210
8. B. Alatas, Chaotic harmony search algorithms. Appl. Math. Comput. **216**, 2687–2699 (2010)
9. B. Alatas, ACROA: artificial chemical reaction optimization algorithm for global optimization. Expert Syst. Appl. **38**, 13170–13180 (2011)
10. O.M. Alia, R. Mandava, The variants of the harmony search algorithm: an overview. Artif. Intell. Rev. **36**, 49–68 (2011)
11. A.R.A. Alsewari, K.Z. Zamli, Design and implementation of a harmony-search-based variable-strength t-way testing strategy with constraints support. Inf. Softw. Technol. **54**, 553–568 (2012)
12. A.R.A. Alsewari, K.Z. Zamli, A harmony search based pairwise sampling strategy for combinatorial testing. Int. J. Phys. Sci. **7**(7), 1062–1072 (2012)
13. M.T. Ameli, M. Shivaie, S. Moslehpour, Transmission network expansion planning based on hybridization model of neural networks and harmony search algorithm. Int. J. Ind. Eng. Comput. **3**, 71–80 (2012)
14. C. Anandaraman, A.V.M. Sankar, R. Natarajan, A new evolutionary algorithm based on bacterial evolution and its applications for scheduling a flexible manufacturing system. Jurnal Teknik Industri **14**(1), 1–12 (2012)
15. A. Askarzadeh, A. Rezazadeh, A grouping-based global harmony search algorithm for modeling of proton exchange membrane fuel cell. Int. J. Hydrogen Energy **36**, 5047–5053 (2011)
16. P. Aungkulanon, P. Luangpaiboon, *Hybridisations of variable neighbourhood search and modified simplex elements to harmony search and shuffled frog leaping algorithms for process optimisations*. Paper presented at the LAENG Transactions on Engineering Technologies, Special Edition of the International MultiConference of Engineers and Computer Scientists (2010)
17. M.T. Ayvaz, Simultaneous determination of aquifer parameters and zone structures with fuzzy C-means clustering and meta-heuristic harmony search algorithm. Adv. Water Resour. **30**, 2326–2338 (2007)
18. A. Bahrololoum, H. Nezamabadi-pour, H. Bahrololoum, M. Saeed, A prototype classifier based on gravitational search algorithm. Appl. Soft Comput. **12**, 819–825 (2012)
19. M. Batty, B. Jiang, Multi-agent simulation: new approaches to exploring space-time dynamics within GIS Working Paper Series, Paper 10. University College London: Centre for Advanced Spatial Analysis (1999)
20. M.A. Behrang, E. Assareh, M. Ghalambaz, M.R. Assari, A.R. Noghrehabadi, Forecasting future oil demand in Iran using GSA (gravitational search algorithm). Energy **36**, 5649–5654 (2011)
21. K.E. Boulding, General systems theory:the skeleton of science. Manag. Sci. **2**(3), 197–208 (1956)
22. A. Chatterjee, G.K. Mahanti, N. Pathak, Comparative performance of gravitational search algorithm and modified particle swarm optimization algorithm for synthesis of thinned scanned concentric ring array antenna. Prog. Electromagn Res. B **25**, 331–348 (2010)
23. H. Chen, S. Li, Z. Tang, Hybrid gravitational search algorithm with random-key encoding scheme combined with simulated annealing. Int. J. Comput. Sci. Netw. Secur. **11**(6), 208–217 (2011)

24. T.Y. Choi, K.J. Dooley, M. Rungtusanatham, Supply networks and complex adaptive systems: control versus emergence. J. Oper. Manag. **19**, 351–366 (2001)
25. R. Damodaram, M.L. Valarmathi, Phishing website detection and optimization using modified bat algorithm. Int. J. Eng. Res. Appl. **2**(1), 870–876 (2012)
26. S. Das, Intelligent market-making in artificial financial markets. Unpublished Master Thesis, Massachusetts Institute of Technology 2003
27. S. Duman, U. Güvenç, Y. Sönmez, N. Yörükeren, Optimal power flow using gravitational search algorithm. Energy Convers. Manag. **59**, 86–95 (2012)
28. M. Dworkis, D. Huang, Genetic algorithms and investment strategy development: Report: 12 May 2008, The Wharton School, University of Pennsylvania 2008
29. M. Eslami, H. Shareef, A. Mohamed, M. Khajehzadeh, Gravitational search algorithm for coordinated design of PSS and TCSC as damping controller. J. Central South Univ. Technol. **19**(4), 923–932 (2012)
30. G.I. Evers, An automatic regrouping mechanism to deal with stagnation in particle swarm optimization. Unpublished Master Thesis, University of Texas-Pan American 2009
31. G.I. Evers, Particle swarm optimization research toolbox documentation: version: 20110515i (2011) www.georgeevers.org/pso_research_toolbox.htm. Accessed 06 June 2013
32. M. Gauci, T.J. Dodd, R. Groß, Why 'GSA: a gravitational search algorithm' is not genuinely based on the law of gravity. Nat. Comput. **11**(4), 719–720 (2012)
33. M. Ghalambaz, A.R. Noghrehabadi, M.A. Behrang, E. Assareh, A. Ghanbarzadeh, N. Hedayat, A hybrid neural network and gravitational search algorithm (HNNGSA) method to solve well known Wessinger's equation. World Acad. Sci. Eng. Technol. **73**, 803–807 (2011)
34. R.L. Goldstone, U. Wilensky, Promoting transfer by grounding complex systems principles. J. Learn. Sci. **17**(4), 465–516 (2008)
35. A. Gorbenko, V. Popov, The force law design of artificial physics optimization for robot anticipation of motion. Adv. Stud. Theor. Phys. **6**(13), 625–628 (2012)
36. T.E. Gorochowski, M.D. Bernardo, C.S. Grierson, Evolving dynamical networks: a formalism for describing complex systems. Complexity **17**, 18–25 (2012)
37. X. Han, X. Chang, A chaotic digital secure communication based on a modified gravitational search algorithm filter. Inf. Sci. **208**, 14–27 (2012)
38. X. Han, X. Chang, Chaotic secure communication based on a gravitational search algorithm filter. Eng. Appl. Artif. Intell. **25**, 766–774 (2012)
39. G. Hartvigsen, A. Kinzing, G. Peterson, Use and analysis of complex adaptive systems in ecosystem science: overview of special section. Ecosystems **1**, 427–430 (1998)
40. A. Hatamlou, S. Abdullah, H. Nezamabadi-pour, A combined approach for clustering based on K-means and gravitational search algorithms. Swarm and Evolutionary Computation **6**, 47–52 (2012)
41. J.H. Holland, *Adaptation in Neural and Artificial Systems* (University of Michigan Press, MI, 1975)
42. J.H. Holland, *Adaptation in Natural and Artificial Systems : An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, 2nd edn. (MIT Press, Cambridge, 1992)
43. J.H. Holland, *Hidden order: how adaptation builds complexity* (Helix Books, Addison-Wesley, New York, 1995)
44. J.H. Holland, Exploring the evolution of complexity in signaling networks. Complexity **7**, 34–45 (2001)
45. J.H. Holland, Complex adaptive systems and spontaneous emergence, in *Complexity and Industrial Clusters*, ed. by A.Q. Curzio, M. Fortis (Physica, Heidelberg, 2002), pp. 25–34
46. J.H. Holland, Studying complex adaptive systems. J. Syst. Sci. Complexity **19**(1), 1–8 (2006)
47. K. Ioannidis, G.C. Sirakoulis, I. Andreadis, Cellular ants: a method to create collision free trajectories for a cooperative robot team. Robot. Auton. Syst. **59**, 113–127 (2011)

48. D. Ivanov, B. Sokolov, The inter-disciplinary modelling of supply chains in the context of collaborative multi-structural cyber-physical networks. J. Manuf. Technol. Manag. **23**(8), 976–997 (2012)
49. M. Kampouridis, Computational intelligence in financial forecasting and agent-based modeling: applications of genetic programming and self-organizing maps. Unpublished Doctoral Thesis, University of Essex (2011)
50. N. Keshavarz, D. Nutbeam, L. Rowling, F. Khavarpour, Schools as social complex adaptive systems: a new way to understand the challenges of introducing the health promoting schools concept. Soc. Sci. Med. **70**, 1467–1474 (2010)
51. M. Khajehzadeh, M. Eslami, Gravitational search algorithm for optimization of retaining structures. Indian J. Sci. Technol. **5**(1), 1821–1827 (2012)
52. M. Khajehzadeh, M.R. Taha, A. El-Shafie, M. Eslami, A modified gravitational search algorithm for slope stability analysis. Eng. Appl. Artif. Intell. **25**(8), 1589–1597 (2012)
53. B. LeBaron, Agent-based computational finance: suggested readings and early research. J. Econ. Dyn. Control **24**, 679–702 (2000)
54. B. LeBaron, Empirical regularities from interacting long- and short-memory investors in an agent-based stock market. IEEE Trans. Evol. Comput. **5**(5), 442–455 (2001)
55. B. LeBaron, W.B. Arthur, R. Palmer, Time series properties of an artificial stock market. J. Econ. Dyn. Control **23**, 1487–1516 (1999)
56. T.A. Lemma,F.B.M. Hashim, *Use of fuzzy systems and bat algorithm for exergy modeling in a gas turbine generator.* Paper presented at the IEEE Colloquium on Humanities, Science and Engineering Research (CHUSER), 5–6 December, Penang, pp. 305–310 (2011)
57. S.A. Levin, Ecosystems and the biosphere as complex adaptive systems. Ecosystems **1**, 431–436 (1998)
58. C. Li, J. Zhou, Parameters identification of hydraulic turbine governing system using improved gravitational search algorithm. Energy Convers. Manag. **52**, 374–381 (2011)
59. C. Li, J. Zhou, J. Xiao, H. Xiao, Parameters identification of chaotic system by chaotic gravitational search algorithm. Chaos, Solitons Fractals **45**, 539–547 (2012)
60. H. Li, *Financial prediction and trading* via *reinforcement learning and soft computing.* Unpublished Doctoral Thesis, University of Missouri-Rolla (2005)
61. P. Li, H. Duan, Path planning of unmanned aerial vehicle based on improved gravitational search algorithm. Sci. China Technol. Sci. **55**(10), 2712–2719 (2012)
62. F.M. Longin, The asymptotic distribution of extreme stock market returns. J. Bus. **69**, 383–408 (1996)
63. E.F.P. da Luz, J.C. Becceneri, H.F. de campos Velho, A new multi-particle collision algorithm for optimization in a high performance environment. J. Comput. Interdisc. Sci. **1**(1), 3–10 (2008)
64. E.F.P. da Luz, J.C. Becceneri, H.F. de campos Velho, *Multiple particle collision algorithm applied to radiative transference and pollutant localization inverse problems.* Paper presented at the IEEE international symposium on parallel and distributed processing workshops and Ph.D. forum (IPDPSW), pp. 347–351 (2011)
65. M.J. Mauboussin, Revisiting market efficiency: the stock market as a complex adaptive system. J. Appl. Corp. Finan. **14**, 47–55 (2002)
66. B. McKelvey, C. Wycisk, M. Hülsmann, Designing an electronic auction market for complex 'smart parts' logistics: options based on LeBaron's computational stock market. Int. J. Prod. Econ. **120**, 476–494 (2009)
67. M.D. Mills-Harris, A. Soylemezoglu, C. Saygin, Adaptive inventory management using RFID data. Int. J. Adv. Manuf. Technol. **32**, 1045–1051 (2007)
68. S. Mirjalili, S.Z.M. Hashim, *A new hybrid PSOGSA algorithm for function optimization.* Paper presented at the proceedings of the international conference on computer and information application (ICCIA), pp. 374–377 (2010)

69. F.S. Mishkin, *The Economics of Money, Banking, and Financial Markets* (The Addison-Wesley, Reading, 2004)
70. L. Monostori, K. Ueda, Design of complex adaptive systems: introduction. Adv. Eng. Inform. **20**, 223–225 (2006)
71. P. Musikapun, P. Pongcharoen, *Solving multi-stage multi-machine multi-product scheduling problem using bat algorithm.* Paper presented at the 2nd international conference on management and artificial intelligence, vol. 35, pp. 98–102 (2012)
72. J.V. Neumann, *Theory of Self-Reproducing Automata* (University of Illinois Press, Urbana, 1966)
73. E.W.T. Ngai, D.C.K. Chau, J.K.L. Poon, A.Y.M. Chan, B.C.M. Chan, W.W.S. Wu, Implementing an RFID-based manufacturing process management system: lessons learned and success factors. J. Eng. Tech. Manage. **29**, 112–130 (2012)
74. T. Niknam, F. Golestaneh, A. Malekpour, Probabilistic energy and operation management of a microgrid containing wind/photovoltai/fuel cell generation and energy storage devices based on point estimate method and self-adaptive gravitational search algorithm. Energy. **43**(1), 427–437 (2012)
75. R.G. Palmer, W.B. Arthur, J.H. Holland, B. LeBaron, An artificial stock market. Artif. Life Robot. **3**, 27–31 (1999)
76. J.P. Papa, A. Pagnin, S.A. Schellini, A. Spadotto, R.C. Guido, M., Ponti, G. Chiachia, A.X. Falcão, *Feature selection through gravitational search algorithm.* Paper presented at the IEEE international conference on acoustics speech (ICASSP), pp. 2052–2055 (2011)
77. S.D. Pathak, J.M. Day, A. Nair, W.J. Sawaya, M.M. Kristal, Complexity and adaptivity in supply networks: building supply network theory using a complex adaptive systems perspective. Decis. Sci. **38**(4), 547–580 (2007)
78. S.D. Pathak, D.M. Dilts, G. Biswas, *Simulating growth dynamics in complex adaptive supply networks.* Paper presented at the 2004 winter simulation conference, pp. 774–782 (2004)
79. P. Rabanal, I. Rodríguez, F. Rubio, Using river formation dynamics to design heuristic algorithms. ed. by C.S. Calude, S.G. Akl, M.J. Dinneen, G. Rozenber, H.T. Wareham , UC 2007, LNCS, vol. 4618 (Springer, Heidelberg, 2007) pp. 163–177
80. P. Rabanal, I. Rodríguez, F. Rubio, *Finding Minimum Spanning/Distances Trees by Using River Formation Dynamics,* vol. 5217, ed. by M. Dorigo, ANTS 2008, LNCS 5217 (Springer, Berlin, 2008a) pp. 60–71
81. P. Rabanal, I. Rodríguez, F. Rubio, *Solving dynamic TSP by using river formation dynamics.* Paper presented at the 4th international conference on natural computation (ICNC), pp. 246–250 (2008b)
82. P. Rabanal, I. Rodríguez, F. Rubio, *Applying river formation dynamics to the Steiner tree problem.* Paper presented at the 9th IEEE international conference on cognitive informatics (ICCI), pp. 704–711 (2010)
83. T. Rambharose, Artificial neural network training add-in for PSO research toolbox. Department of Computing & Information Technology, The University of the West Indies, St. Augustine (2010), http://www.tricia-rambharose.com. Accessed 06 June 2013
84. E. Rashedi, H. Nezamabadi-pour, S. Saryazdi, GSA: a gravitational search algorithm. Inf. Sci. **179**, 2232–2248 (2009)
85. E. Rashedi, H. Nezamabadi-pour, S. Saryazdi, BGSA: binary gravitational search algorithm. Nat. Comput. **9**(3), 727–745 (2010)
86. E. Rashedi, H. Nezamabadi-pour, S. Saryazdi, Filter modeling using gravitational search algorithm. Eng. Appl. Artif. Intell. **24**, 117–122 (2011)
87. P.K. Roy, B. Mandal, K. Bhattacharya, Gravitational search algorithm based optimal reactive power dispatch for voltage stability enhancement. Electr. Power Compon. Syst. **40**, 956–976 (2012)
88. B. Rundh, Radio frequency identification (RFID): invaluable technology or a new obstacle in the marketing process? Mark. Intell. Planning **26**(1), 97–114 (2008)

89. W.F. Sacco, C.R.E. de Oliveira, *A new stochastic optimization algorithm based on a particle collision metaheuristic*. Paper presented at the 6th World Congresses of Structural and Multidisciplinary Optimization (Rio de Janeiro, 30 May–03 June 2005) pp. 1–6
90. S. Sarafrazi, H. Nezamabadi-pour, S. Saryazdi, Disruption: a new operator in gravitational search algorithm. Scientia Iranica D **18**(3), 539–548 (2011)
91. B. Shaw, V. Mukherjee, S.P. Ghoshal, A novel opposition-based gravitational search algorithm for combined economic and emission dispatch problems of power systems. Electr. Power Energ. Syst. **35**, 21–33 (2012)
92. S. Soni, Applications of ANNs in the stock market prediction: a survey. Int. J. Comput. Sci. Eng. Technol. **2**(3), 71–83 (2010)
93. H.S. Sudhira, *Integration of agent-based and cellular automata models for simulating urban sprawl.* Unpublished Master Thesis, International Institute for Geo-Information Science and Earth Observation & Department of Space, Indian Institute of Remote Sensing, National Remote Sensing Agency (NRSA) (Enschede, Dehradun, 2004)
94. N. Suhadolnik, J. Galimberti, S.D. Silva, Robot traders can prevent extreme events in complex stock markets. Physica A **389**, 5182–5192 (2010)
95. A. Surana, S. Kumara, M. Greaves, U.N. Raghavan, Supply-chain networks: a complex adaptive systems perspective. Int. J. Prod. Res. **43**(20), 4235–4365 (2005)
96. J.M. Swaminathan, S.F. Smith, N.M. Sadeh, Modeling supply chain dynamics: a multiagent approach. Decis. Sci. **29**(3), 607–632 (1998)
97. M. Taherdangkoo, M.H. Shirzadi, M.H. Bagheri, A novel meta-heuristic algorithm for numerical function optimization: blind, naked mole-rats (BNMR) algorithm. Sci. Res. Essays **7**(41), 3566–3583 (2012)
98. J. Tan, H.J. Wen, N. Awad, Health care and services delivery systems as complex adaptive systems. Commun. ACM **48**(5), 36–44 (2005)
99. L.D. Thurston, Jacksonville to construct first refrigerated crossdock. Caribbean Bus. **36**(40), 41 (2008)
100. P. Wang, Y. Cheng, Relief supplies scheduling based on bean optimization algorithm. Econ. Res. Guide **8**, 252–253 (2010)
101. R.A. Watson, C.L. Buckley, R. Mills, Optimization in self-modeling complex adaptive systems. Complexity **16**, 17–26 (2011)
102. Y.-M. Wei, S.-J. Ying, Y. Fan, B.-H. Wang, The cellular automaton model of investment behavior in the stock market. Phys. A **325**, 507–516 (2003)
103. U. Wilensky, *NetLogo (Version 4.1) center for connected Learning and Computer-Based Modeling* http://ccl.northwestern.edu/netlogo/ (Northwestern University, Evanston, 1999)
104. Y. Wu, A dual-response strategy for global logistics under uncertainty: a case study of a third-party logistics company. Int. Trans. Oper. Res. **19**(3), 397–419 (2012)
105. C. Wycisk, B. McKelvey, M. Hülsmann, "Smart parts" supply networks as complex adaptive systems: analysis and implications. Int. J. Phys. Distrib. Logist. Manag. **38**(2), 108–125 (2008)
106. L. Xie, J. Zeng, R.A. Formato, Convergence analysis and performance of the extended artificial physics optimization algorithm. Appl. Math. Comput. **218**, 4000–4011 (2011)
107. B. Xing, W.-J. Gao, *Computational Intelligence in Remanufacturing* (IGI Global, Hershey, 2014) ISBN 978-1-4666-4908-8
108. B. Xing, and W.-J. Gao, *Innovative Computational Intelligence: A Rough Guide to 134 Clever Algorithms* (Springer, Cham, 2014) ISBN 978-3-319-03403-4
109. S.-D. Yang, Y.-L. Yi, Z.-Y. Shan, Gbest-guided artificial chemical reaction algorithm for global numerical optimization. Procedia Eng. **24**, 197–201 (2011)
110. X.-S. Yang, A.H. Gandomi, Bat algorithm: a novel approach for global engineering optimization. Eng. Comput. **29**(5), 464–483 (2012)
111. C.Y. Yi, E.W.T. Ngai, K.-L. Moon, Supply chain flexibility in an uncertain environment: exploratory findings from five case studies. Supply Chain Manag. Int. J. **16**(4), 271–283 (2011)

112. M. Yin, Y. Hu, F. Yang, X. Li, W. Gu, A novel hybrid K-harmonic means and gravitational search algorithm approach for clustering. Expert Syst. Appl. **38**, 9319–9324 (2011)
113. X. Zhang, K. Jiang, H. Wang, W. Li, B. Sun, *An Improved Bean Optimization Algorithm for Solving TSP*, vol. 7331, ed. by Y. Tan, Y. Shi, Z. Ji, ICSI 2012, Part I, LNCS 7331 (Springer, Berlin, 2012), pp. 261–267
114. X. Zhang, B. Sun, T. Mei, R. Wang, *Post-disaster restoration based on fuzzy preference relation and bean optimization algorithm.* Paper presented at the IEEE Youth Conference onInformation Computing and Telecommunications (YC-ICT), (28–30 Nov 2010), pp. 271–274
115. Z.-N. Zhang, Z.-L. Liu, Y. Chen, Y.-B. Xie, Knowledge flow in engineering design: an ontological framework. Proc. Inst. Mech. Eng. [C] J. Mech. Eng. Sci. **227**(4), 760–770 (2013)
116. W. Zhou, S. Piramuthu, Remanufacturing with RFID item-level information: optimization, waste reduction and quality improvement. Int. J. Prod. Econ. **145**(2), 647–657 (2013)
117. B. Zibanezhad, K. Yamanifar, R.S. Sadjady, Y. Rastegari, Applying gravitational search algorithm in the QoS-based Web service selection problem. J. Zhejiang Univ. Sci. C (Comput. Electron.), **12**(9), 730–742 (2011)

Optimized Reconfigurable Autopilot Design for an Aerospace CPS

Arsalan H. Khan, Zeashan H. Khan and Salman H. Khan

Abstract A modular flight control strategy is presented here to demonstrate the improved command tracking performance with fault tolerance and reconfiguration capabilities. The modular control design process consists of inner and outer loop design concept, where outer baseline controller feedback loop ensures the stability and robustness and inner reconfigurable design is responsible for the fault-tolerance against actuator faults/failures. This guarantees augmented autonomy and intelligence on board aircraft for real time decision and fault tolerant control. Requirements for aerospace cyber physical systems (ACPS) and software are far more stringent than those found in industrial automation systems. The results shows that fault tolerant aspect is mandatory for ACPS, that must support real time behavior and also requires ultra-high reliability as many systems or/sub-systems are safety critical and require certification.

1 Introduction

A fly-by-wire aircraft has a modern avionics architecture integrating the physical and the communication layer in the aircraft's cockpit. Thus, a fly-by-wire digital autopilot system is a distributed control system with a number of on-board computers, sensors and redundant actuators to provide maximum survivability in the case of an accident by ensuring physical as well as analytical redundancy.

A. H. Khan (✉)
Northwestern Polytechnical University (NPU), Xi'an, China
e-mail: arsalan556@yahoo.com

Z. H. Khan
National University of Sciences and Technology (NUST), Islamabad, Pakistan

S. H. Khan
University of Western Australia (UWA), Perth, Australia

Z. H. Khan et al. (eds.), *Computational Intelligence for Decision Support in Cyber-Physical Systems*, Studies in Computational Intelligence 540,
DOI: 10.1007/978-981-4585-36-1_13,

The main emphasis is now on digital computing with the use of inertial motion and air stream sensor units; the direct mechanical linkages between the cockpit controls and the control surfaces have been removed and replaced with electrical signaling with direct motion commands, hence the term 'fly-by-wire'. There is also some development to incorporate fiber optic communication in the onboard aircraft hence named 'fly-by-light'.

As compared to the centralized control strategy in which a single controller is employed to regulate the output 'y' according to the commanded reference input 'r', modular strategy has more design freedom for control engineers. Several centralized strategies are available in literature [1–5], where in [1] a single controller is designed for speed regulation of automotive engine and in [5] a centralized robust controller is used for UAV flight path tracking. It is sometimes quite difficult for the designers to determine a single optimal controller for achieving multiple objectives in the complete complex system where the inputs are interdependent. Due to the design complexity of centralized control approach by considering all the system dynamics, it is sometimes not possible and effective to choose this architecture. The applicability of one large centralized control system is also restricted because of the high computational power requirement and time consumption for reconfiguration in case of fault and failure scenarios. Advantages of centralized strategies include the absence of communication delays, information loss and other integration issues between subsystem controllers, guaranteed stability for designed set-point, and rigorous theoretical proofs.

An example of one of the module in aerospace CPS is shown in Fig. 1. The air data system is designed to provide high integrity information; for example, the arrangement in figure above might provide triplex airspeed information necessary which plays a vital role in safe flight. In practice, the quality and integrity of the air data will depend on the capabilities and locations of the individual sensors. For the arrangement shown in Fig. 1, 'a' is a static probe, and 'b', 'c' and 'd' are multihole probes used to resolve local flow angles from dynamic pressure data. The air-data information is complemented with information from the aircraft's inertial sensors.

Modular control approach is now becoming a standard for safety-critical systems where several small modular control loops are present, each dedicated to handle a subsystem and its dynamics. The advantages of modular strategy are the feasibility of modification in case of sub-system up-gradation, as the subsystem constraints can be easily handled, efficient use of parallel computing capabilities for fast execution, easy identification of faults/failures, highly recommended by control system designers for large problems, and accessibility of specialized efficient optimal dedicated controllers and algorithms for improved performance in nominal and in fault cases.

We focus on the aerospace industry challenges and procedures to build high confidence aerospace cyber physical systems (Aerospace CPSs) which require:

1. The algorithms and protocols to be designed should come with guarantees of correctness to ensure maximum survivability. In particular, it should be possible to formally prove that under reasonable assumptions on the environment

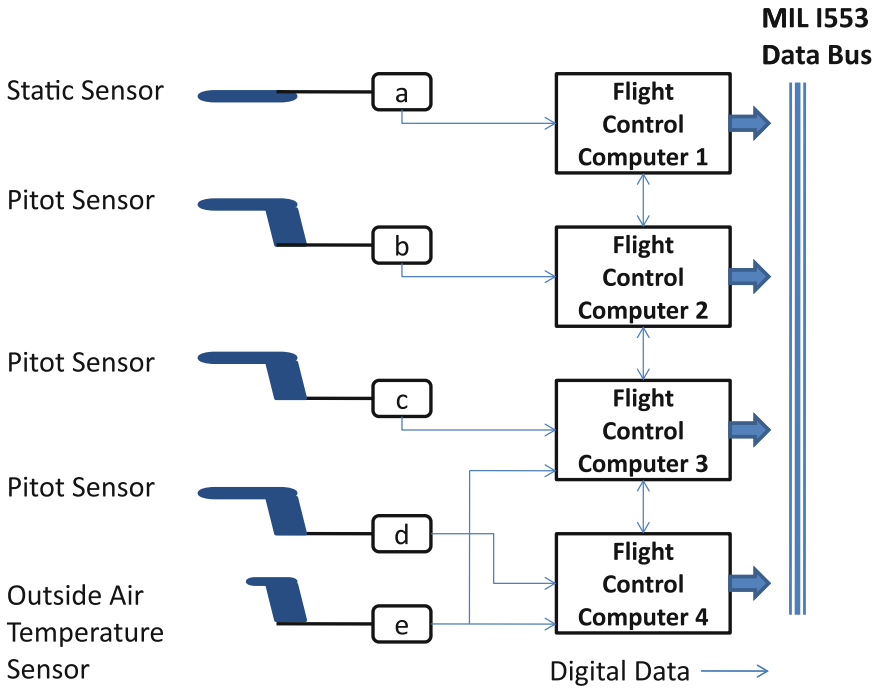


Fig. 1 Air data system architecture of a fighter aircraft

(in terms of communication networks, computation resources, and the relevant physical processes); the algorithms are guaranteed to achieve a certain level of performance.

2. The overall system performance should not be fragile with respect to small perturbations to the environment nor, to the extent possible, failures in the embedded systems that monitor or control this environment. In case, it is not possible to ensure total performance, a graceful performance degradation and execution of emergency protocol must be ensured.
3. The cross-domain elements of CPSs including electrical, mechanical and hydraulic system, core control and monitoring system, software, and communication network should be modeled in a unified way.
4. Calibration, testing, and diagnosis of these complex systems should be very reliable and accurate.

2 Modular Fault Tolerant Control

The modular approach consists of inner loop and outer loop configuration as presented in Fig. 2, has dominant performance especially for over-actuated systems as demonstrated in [6–10]. The concept of virtual control command v is

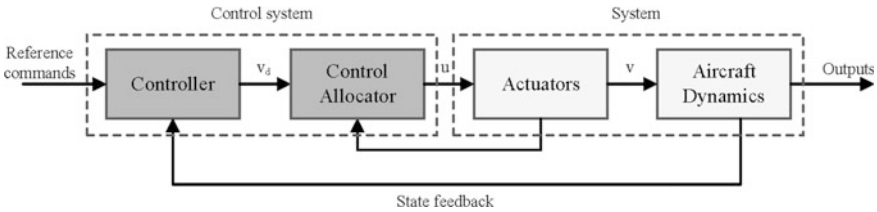


Fig. 2 Modular (inner/outer loop) controller configuration

presented in [6–10], which represents a demanded force, moment, or generalized objective that must be achieved through the optimal combination of all the available redundant actuators. In flight control systems the virtual command is generally consist of roll, pitch, and yaw moments (L, M, N). The actual control commands are the control surface deflections, brake pressures, and engine throttle etc. which produce the combine effects of the virtual control commands. The baseline controller in outer loop is responsible to produce virtual commands which are sufficient to drive the plant dynamics (the aircraft dynamics, in this case). As illustrated in Fig. 2 a control allocation module is introduced for optimized and effective distribution of baseline controller generated virtual commands between the redundant actuators.

Control allocation is a very useful methodology in active fault-tolerant control for over-actuated systems. The idea of control allocation is the separation of actuator selection and command tracking task in the control design. The concept of control allocation is an active research topic in marine, aerospace and automotive vehicle control [7, 11–24]. To control the dynamic position of marine vehicles, thrusters, propellers and rudders are used to produce translational forces and yawing moment, in order to keep the heading in the desired direction [11–15]. In the advanced aerospace applications actuator redundancy is available for improved maneuverability, reliability, and safety. The primary and secondary control surfaces include aileron, elevator, rudder, and canards, spoilers, flaps, and thrust vectoring vanes respectively, to produce the net torque acting on the vehicle for controlling its motion [7, 16–20]. In automotive vehicle control, the net yawing moment depends on the individual brake forces produced by each wheel [21–25]. Optimal control allocation for flapping wing micro air vehicle control, walking robots, and large-scale air-jet actuators array are some of the example to improve the performance and efficiency of over-actuated mechanical systems [26–29]. The modular flight controller design for over-actuated systems is performed in the following two steps.

1. Design a robust baseline control law specifying total control effort to be produced (moments, forces, etc.)
2. Design a control allocator that distribute the total demanded effort onto the actuators according to their specifications (control surfaces deflections, thrust forces, etc.)

3 Real Time Control Allocation

The online computation requirements and practical implementation issues of control allocators in demanding flight control systems are of crucial concern. Control allocation methods based on weighted pseudo inverse (WPI) has been used in real aircraft flight control system where large number of similar control surfaces is available, such as Boeing 747. Cascaded pseudo inverse (CGI) method is applied in X-35 aircraft for control allocation [30]. Davidson et al. [31], presented an extended version of WPI method for desktop and real-time piloted simulation of tailless aircraft. The applicability of pseudo inverse based CA solution because of its fast execution draws researchers to compare their proposed real-time solutions with this strategy [32–35]. In [34], Durham presented BESA implementation results and compare it with the generalized inverse (GI) solutions. It was found that BESA approach for three objective control allocation is slower than GI methods but faster than CGI method. Bodson [36] compared his proposed solution based on the Simplex solution to a LP problem with RPI method. It was found that execution time of Bodson direction preserving method is five time more time consuming than RPI. In [37], Harkegard suggested least squares based control allocation solution and found that it is computationally similar to CGI method but not as accurate as CGI for feasible commands. However, a considerable drawback of pseudo inverse based solutions is their incapability to achieve the entire feasible commanded objectives because of low AMS volume as compare to direct control allocation AMS. In [38, 39], authors have suggested global optimization based solutions to maximize the AMS volume of pseudo-inverse based methods.

Optimization based CA strategies are very promising in the accuracy of solution but the computational load of these algorithms is quite exhaustive. Recently, several optimization based control allocation strategies are considered for real-time implementation because of increase in computational capabilities [21, 40–43]. In [40], Bolender et al. presented a mixed-integer linear programming (MILP) based control allocation solution for reusable launch vehicle considering piecewise linear models. The validation and verification of MILP solvers is very difficult especially for use in safety critical real-time applications, which makes use of this strategy limited. Optimization based control allocation strategies are applied in ground vehicles using real-time optimizing program [21, 43]. In automotive vehicles, fast optimization solvers can be used for control allocation as considerable amount of time is available in on-ground moving vehicles as compared to high speed fighter aircrafts. Some real-time implementation and simulation results of optimization based control allocation strategies using fast optimization solvers will be presented in next few sections.

4 Reconfigurable Modular FCS Design for ADMIRE Benchmark

To demonstrate the performance of reconfigurable modular control strategy a linearized ADMIRE aircraft model is selected at low speed flight condition i.e., Mach 0.22 at an altitude of 3000 m. At low speed, the efficiency of the control surfaces is poor so better evaluation can be performed for the case of different actuator faults. The linearized dynamics of an aircraft at a trim condition is represented in state-space form as:

$$\begin{aligned} \dot{x} &= Ax(t) + Bu(t) \\ y &= Cx(t) + Du(t) \end{aligned} \tag{1}$$

where, $A \in R^{nxn}$, $B \in R^{nxm}$, $C \in R^{pxn}$ and $D \in R^{pxm}$ are respectively the state, the input control, the output and feed-through matrices, $x \in R^{n}$ is the system state vector; $u \in R^{m}$ is the control input vector and $y \in R^{p}$ is the system output vector to be controlled in an optimal way. Here, feed-through matrix D is a null matrix, and all states are measurable and the system is full-state feedback system. Now, for incorporating actuator faults or failures, we introduce a diagonal gain matrix $K_e \in R^{mxm}$ in Eq. 1.

The post-fault estimated state-space form becomes

$$\dot{x}(t) = Ax(t) + BK_e u(t) \tag{2}$$

With,

$$K_e = \begin{bmatrix} k_1 & 0 & \cdots & 0 \\ 0 & k_2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & k_m \end{bmatrix}$$

where, $k_i = 1, i = 1 \ldots m$ corresponds the perfect condition of ith control actuator, $0 < k_i < 1$ indicates the fault presence and $k_i = 0$ denotes the total failure or loss of control of the ith control actuator. In order to introduce virtual control command concept [32], for redundant actuators, we assume that the rank $(B) = l < m$. So, the control matrix B can be factorized as

$$B = B_v.B_e \tag{3}$$

where, $B_v \in R^{nxl}$ and $B_e \in R^{lxm}$ are respectively, the virtual control and control effectiveness matrices. The alternate state equation form of Eq. 1 can be written as

$$x(t) = Ax(t) + B_v v(t) \tag{4}$$

$$v(t) = B_e u(t) \tag{5}$$

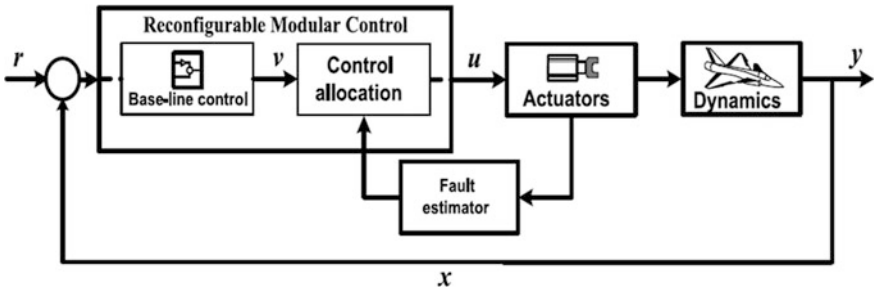


Fig. 3 Reconfigurable modular flight control with post-fault estimator

where $v \in R^l$ is the virtual control command input, known as the total control effort produced by the actuators and commanded by the base-line controller. For the present work, $k = p$ is considered, i.e., the number of virtual control command (v) is equal to the number of outputs (y) to be controlled. In advance aircraft systems, the actuator dynamics are much faster than the aircraft dynamics. So, for the control allocation process, we consider a linear relationship between constrained control command (u) and virtual control command (v) as in Eq. 5. The actuators control command u(t) is limited by

$$u_{min} \leq u \leq u_{max}$$

where, u_{min} and u_{max} are the lower and upper position deflection limits of physical actuators. The reconfigurable modular structure of FCS with control allocator is shown in Fig. 3. Typically, the virtual control command v is a 3-dimensional vector consisting of rolling moment (C_l), pitching moment (C_m) and yawing moment (C_n); while u represents the respective actuator commanded positions. Reliable linear quadratic (LQ) control with control allocator is designed here. Linear programming (LP) based control allocation strategy is used to solve the following cost function

$$\max_{u,a} \quad a$$
$$subject \;\; to \;\; Bu = av$$
$$u_{min} \leq u \leq u_{max}$$

where, 'a' is the scaling factor. The simulation is performed in non-real-time environment using Matlab 7.13. From Eq. 1, following are the state matrix and input control matrices at trim flight condition:

$$A = \begin{bmatrix} -0.5432 & 0.0137 & 0 & 0.9778 & 0 \\ 0 & -0.1179 & 0.2215 & 0 & -0.9661 \\ 0 & -10.5130 & -0.9968 & 0 & 0.6176 \\ 2.6221 & -0.0030 & 0 & -0.5057 & 0 \\ 0 & 0.7076 & -0.0939 & 0 & -0.2127 \end{bmatrix}$$

$$B = \begin{bmatrix} 0.0035 & 0.0035 & -0.0318 & -0.0548 & -0.0548 & -0.0318 & 0.0004 \\ -0.0063 & 0.0063 & 0.0024 & 0.0095 & -0.0095 & -0.0024 & 0.0287 \\ 0.6013 & -0.6013 & -2.2849 & -1.9574 & 1.9574 & 2.2849 & 1.4871 \\ 0.8266 & 0.8266 & -0.4628 & -0.8107 & -0.8107 & -0.4628 & 0.0024 \\ -0.2615 & 0.2615 & -0.0944 & -0.1861 & 0.1861 & 0.0944 & -0.8823 \end{bmatrix}$$

The command input (reference input) consist of angle of attack α (rad), sideslip β (rad), and roll rate p (rad/sec), given as:

$$y = [\alpha \ \beta \ p]^T$$

whereas the state vector consist of pitch rate q (rad/sec) and yaw rate r (rad/sec) with commanded inputs as:

$$x = [\alpha \ \beta \ p \ q \ r]^T$$

There are seven control surfaces $[\delta_{lc}, \delta_{rc}, \delta_{roe}, \delta_{rie}, \delta_{lie}, \delta_{loe}, \delta_r]^T$ representing the left and right canards, right inner and outer elevons, left inner and outer elevons, and the rudder respectively. Position constraints are considered in CA strategy with the following limits:

$$\begin{aligned} u_{min} &= [-55 \quad -55 \quad -30 \quad -30 \quad -30 \quad -30 \quad -30]^T (\text{deg}) \\ u_{max} &= [25 \quad 25 \quad 30 \quad 30 \quad 30 \quad 30 \quad 30]^T (\text{deg}) \end{aligned}$$

The approximate model with control allocator has the following effectiveness matrix, where $B = B_v.B_e$ and $B_v = [0_{2\times3} \ I_{3\times3}]^T$

$$B_e = \begin{bmatrix} 0.6013 & -0.6013 & -2.2849 & -1.9574 & 1.9574 & 2.2849 & 1.4871 \\ 0.8266 & 0.8266 & -0.4628 & -0.8107 & -0.8107 & -0.4628 & 0.0024 \\ -0.2615 & 0.2615 & -0.0944 & -0.1861 & 0.1861 & 0.0944 & -0.8823 \end{bmatrix}$$

Open-loop characteristics of the aircraft at this trim condition are given in Table 1. The system is unstable because of the unstable right half plane pole and insufficient natural frequency. Closed-loop modular controller is suggested for this unstable system with LQ optimal control and LP based control allocation. The closed-loop characteristics are given in Table 2 and the response is shown in Fig. 4.

Three different fault scenarios are presented in Figs. 5, 6 and 7 respectively. In Fig. 5, a canard total failure is considered, where the canard is unable to deflect according to commanded deflection and stuck at zero position. Due to the available redundancy in ADMIRE aircraft, CA algorithm redistributes the control efforts to healthy elevons in tracking the commanded α-roll maneuver.

Figure 6, shows a 50 % fault in canard effectivity at 2 s, which causes an overshoot in pitch tracking. By using CA strategy to redistribute the control effort in healthy elevons for compensating the 50 % loss in canard, a better pitch

Table 1 Open-loop system characteristics at Mach = 0.22 and Alt = 3000 m

States $x(t)$	Eigenvalues	Damping ζ	Natural frequency w_n(rad/sec)
Angle of attack α	−2.13	1.0	2.13
Sideslip β	1.08	−1.0	1.08
Roll rate p	−0.318 + 1.70i	0.38	1.73
Pitch rate q	−0.318 − 1.70i	0.38	1.73
Yaw rate r	−0.692	1.0	0.692

Table 2 Closed-loop system characteristics at Mach = 0.22 and Alt = 3000 m

States $x(t)$	Eigenvalues	Damping ζ	Natural frequency w_n(rad/sec)
Angle of attack α	−1.09	1.0	1.09
Sideslip β	−1.15 + 2.04i	0.49	2.34
Roll rate p	−1.15 − 2.04i	0.49	2.34
Pitch rate q	−2.36	1.0	2.36
Yaw rate r	−2.23	1.0	2.23

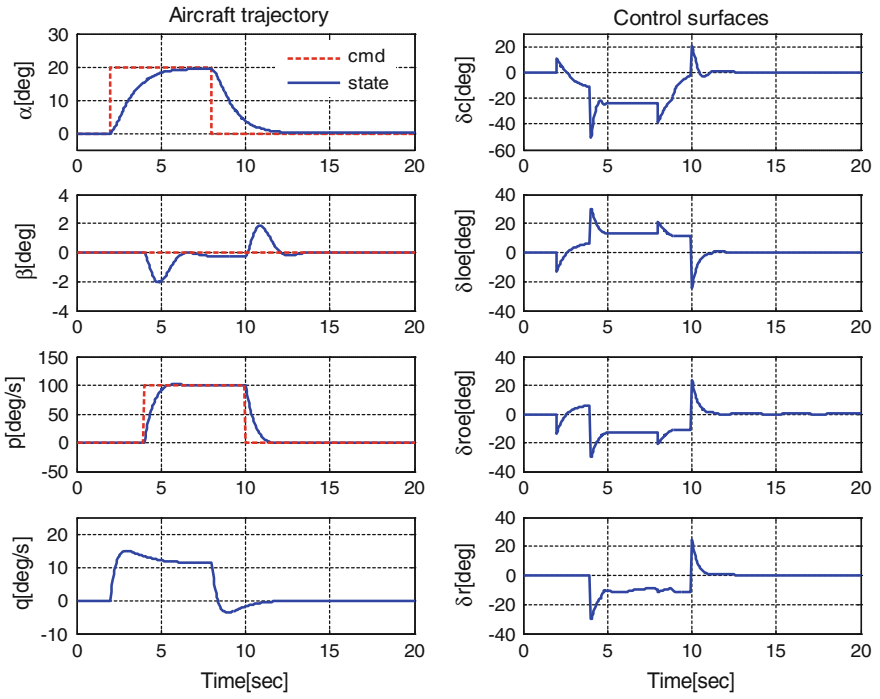


Fig. 4 Aircraft normal flight commanded trajectories and actual trajectories

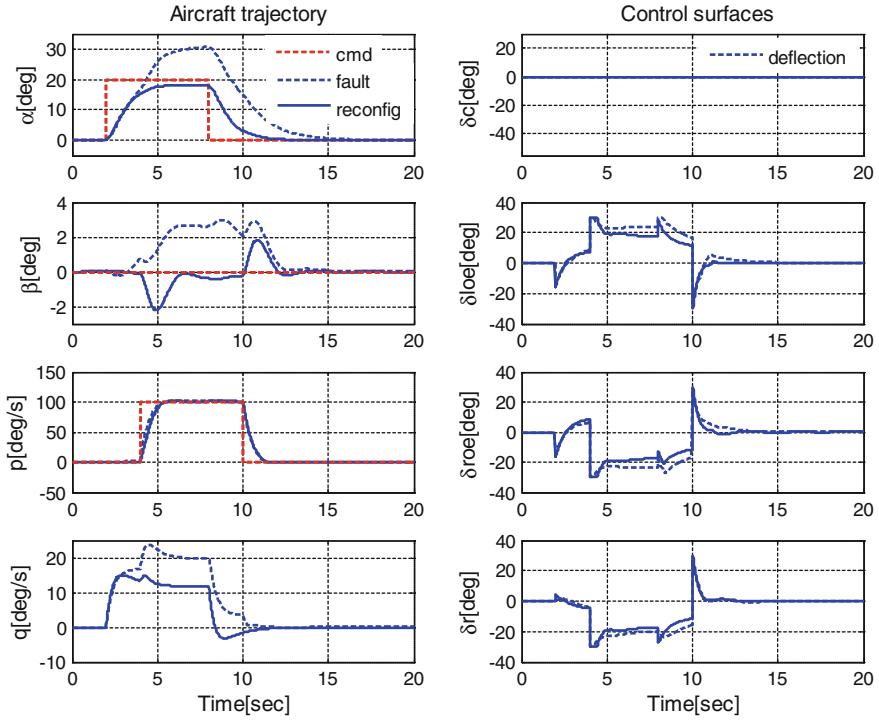


Fig. 5 Canard total failure states and actuators deflections

tracking is achieved as compared to total canard failure case in Fig. 5. In Fig. 7 a 40 % fault is introduced in the left elevon, which causes a severe overshoot in α tracking. Due to the importance of elevon control surface in pitch and roll commands tracking, it is not possible to fully compensate left elevon fault using canard and right elevon. A considerable steady-state error is observed in following α command, whereas p and β tracking remains nominal.

5 Optimization of Reconfigurable Control

Optimization algorithms for modular control strategy are presented here for improving the performance of optimal reconfigurable control design. Two of the most powerful and broadly applicable stochastic search and global optimization techniques each belongs to evolutionary computation and swarm intelligence fields are compared for feedback control loop design optimization [37, 38]. Both evolutionary computation and swarm intelligence are the paradigms of computational intelligence (CI) which relates artificial intelligence (AI) to biological or naturally

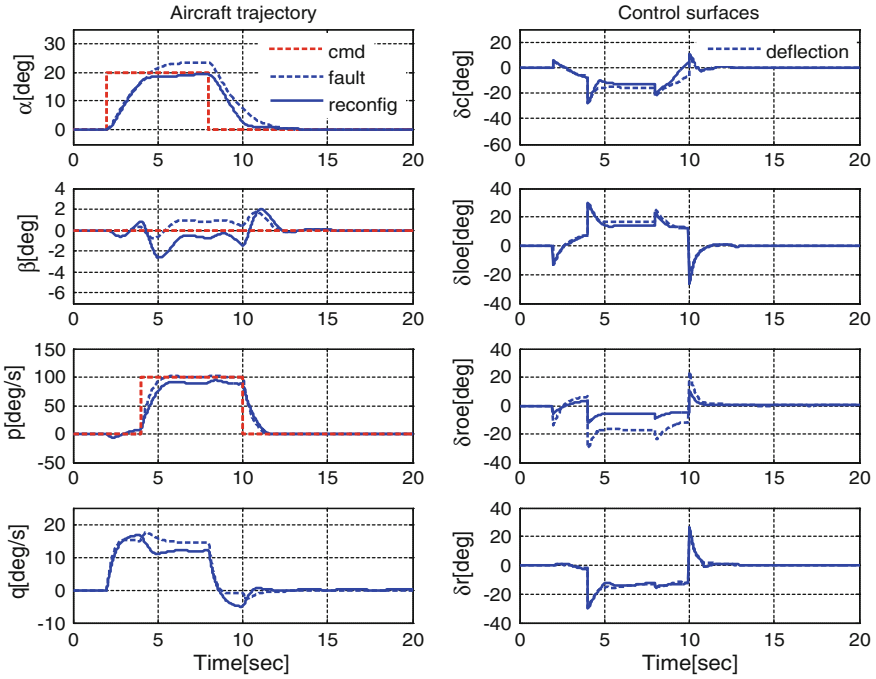


Fig. 6 Canard fault 50 % at time = 2 s: states and actuators deflections

occurring system. Brief description of particle swarm optimization (PSO) related to flight controller design is presented and demonstrated for ADMIRE aircraft in next section.

6 Particle Swarm Optimization

Particle swarm optimization (PSO) is a relatively new stochastic optimization technique, introduced by Kennedy and Eberhart [43, 44] in 1995. Comparatively, it is a computationally efficient optimization method and quite simple than the genetic algorithm (GA). Like GA, PSO is also a population based intelligent optimization technique inspired by the idea of fish schooling, mosquitoes swarming and birds flocking.

In the PSO algorithm, the swarm of particles is initially placed at random positions in problem search-space with given velocities in search of optima. The directions of the movement and the velocities of the particles are then gradually changed according to their own historical best position and other neighborhood particle's best position for searching better positions. Both the particle's best and the neighborhood's best positions are derived according to the given objective

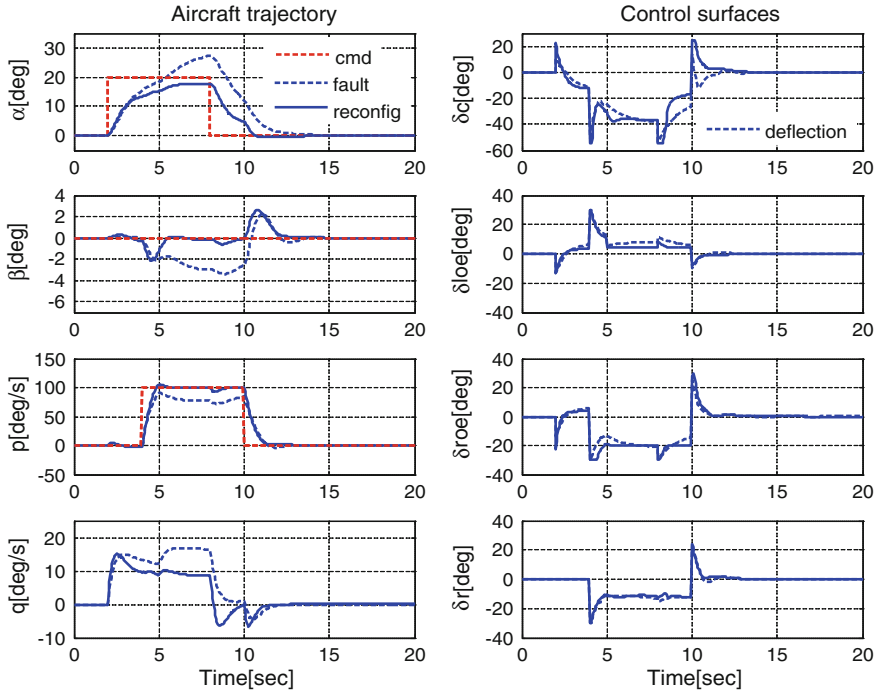


Fig. 7 Left elevon 40 % loss in effectively at time = 5 s: states and actuators deflections

fitness function [43]. PSO has been widely used in state estimation, optimal power flow, antennas design, feedback controller tuning, system identification, and intelligent control [38, 45–48].

7 Working Principle of PSO

PSO is a metaheuristic algorithm like GA, which is based on particles (same as chromosomes in GA). However, unlike the GA, PSO particles move in search of optimum solution using three D-dimensional vectors instead of evolution operators. Where, 'D' is the dimensionality of the problem space, where each particle has three parameters i.e., the current position $\vec{x}_{id}$, the previous best position $\vec{p}_{id}$, and the velocity $\vec{v}_{id}$.

The randomly initialized PSO particles are real value individuals, have their own current position $\vec{x}_{id}$ may be viewed as coordinates describing their location in space. At each step of the PSO algorithm, the current position is evaluated as a solution for optimization. If that position is better than any that has been found in previous steps so far, then the location information is stored in the second vector, $\vec{p}_{id}$. The value of the best function result achieved so far is stored in a variable,

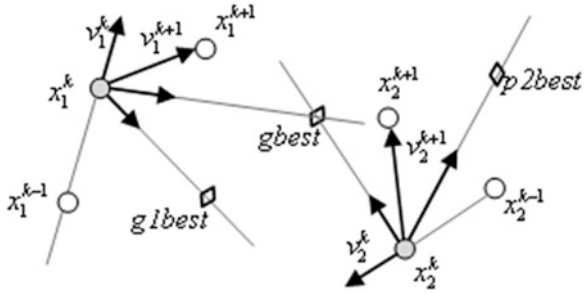


Fig. 8 Particles movement in swarm

p_{ibest} (for "previous best") for evaluation on later steps. New better position is attained by adding $\vec{v}_{id}$ vector to $\vec{x}_{id}$ in search of better solution for objective function and up-gradation of $\vec{p}_{id}$ and p_{ibest}. The velocity vector $\vec{v}_i$ is an adjustable variable, which can be considered as the step size. In moving towards the better position (solution) particles communicate with each other in such a way that their motion is affected by the best solution found by any neighborhood particle. Figure 8 illustrates the movement of particles in 2-dimensional space.

The PSO algorithm execution flow for designing an optimal modular flight controller is shown in Fig. 9. The algorithm proceeds as

1. Initially, a randomly distributed population of m particles is generated in D-dimensional search space with random positions $\vec{x}_{id}\ (x_{i1}, x_{i2}, \ldots, x_{iD})$ and velocities $\vec{v}_{id}\ (v_{i1}, v_{i2}, \ldots, v_{iD})$.
2. ***loop***
3. For each generated particle which signifies a potential solution, evaluate the desired optimization fitness function in D variables.
4. Compare the particle's fitness evaluation with its p_{ibest}. If the current value is better than the p_{ibest}, then update p_{ibest} with current value, and $\vec{p}_{id}$ equal to the current location $\vec{x}_{id}$.
5. Identify the neighborhood particle which has the best success so far, and keep track of its position denoted by $\vec{p}_{gd}\ \left(p_{g1}, p_{g2}, \ldots, p_{gD}\right)$.
6. Calculate a velocity vector for each ith particle using following expression at (k + 1)st iteration

$$\vec{v}_{id}^{k+1} = w\,\vec{v}_{id}^{k} + c_1 r_1\,(\vec{p}_{id} - \vec{x}_{id}^{k}) + c_2\,r_2\,(\vec{p}_{gd} - \vec{x}_{id}^{k}) \tag{6}$$

7. Using the updated velocities and previous position, each particle changes its position as

$$\vec{x}_{id}^{k+1} = \vec{x}_{id}^{k} + \vec{v}_{id}^{k+1} \tag{7}$$

8. If the criterion is met (usually the designated fitness or maximum number of iterations), exit loop.
9. end ***loop***

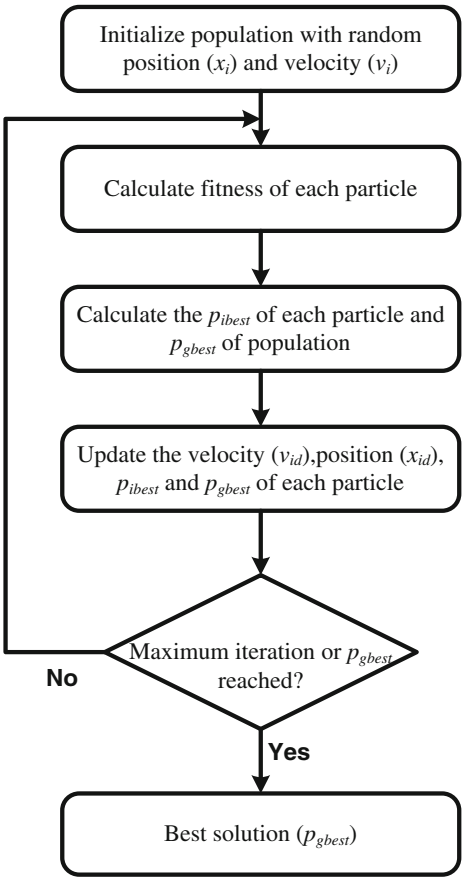


Fig. 9 Flow chart PSO algorithm

In Eq. 6, c_1 and c_2 are the two positive constant parameters known as learning factors, and w is a positive constant called inertia of the particle which are dependent on the optimization problem. Whereas r_1 and r_2 are the uniformly distributed random numbers between $[0, 1]$, and $d = 1, 2, \ldots, D$.

8 PSO Design Considerations

Only three parameters, which include two trust parameters/learning factors (c_1, c_2) and the inertia of particles (w) that need to be design carefully in PSO algorithm as compare to several complex parameters in GAs. The initial random distribution of particles is also not as important as in GAs. The reason is that the swarm changes dynamically throughout the optimization process until an optimum solution is found. However, the size of swarm is important and it is often set on the basis of dimensionality and intuitive knowledge of the problem.

9 Parameters

Inertial weight: Introduction of inertial weight parameter w was motivated by the desire to better control the search objective and suggested by Shi and Eberhart [49]. There are several strategies reported by researchers for adjusting the value of w [50–52]. In [51], the value of w is decreased by fraction if no improvement has been achieved during predefined number of design iterations. Whereas, in [50] a random uniform distribution between $[0.5, 1]$ for w is used as an effective design approach.

Generally, inertia weight w is initialized by the setting values between [0.4, 1]. Relatively large value (e.g., 1) correspond to a system where particles move fast in low viscosity medium and perform extensive exploration in problem space. The small value (e.g., 0.4) correspond to a system where particles move slow and may converge to local optima. Here, for flight control design problem, we used $w = 0.5$–1 to better explore the design problem for global optima. The initialization of swarm with inertia value greater than 1 makes the swarm unstable and it is necessary to reduce it significantly to bring particles in stable region.

Learning factors: The trust parameters c_1 and c_2 control the convergence of the particles and allow a systematic method for preventing explosion and ensuring convergence. In [53], Kennedy et al., suggested $c_1 = c_2 = 2$, such that the mean of the stochastic multipliers in Eq. 6 is equal to 1. Equal values of c_1 and c_2 means that particle has equal trust on itself as on the swarm. Here we use $c_2 = 2.2$ and $c_1 = 1.8$ to make each particle puts little more trust on swarm as compare to itself.

10 PSO for Modular Flight Control

Swarm intelligence based particle swarm optimization technique is used to improve the modular flight controller response through optimization of CA and LQ control.

11 Control Allocation Optimization

Weighted generalized inverse (WGI) based control allocation technique is used here because of its computation efficiency. The allocation efficiency of generalized inverses (GIs) is low and it often fails to allocate feasible control command. There are several GI solutions available for underdetermine system of equations. In flight control system, where multiple control surfaces are available to distribute the three conventional control commands (roll, pitch, yaw), generalized inverses are employed for effective distribution according to the actuators capabilities. Here, an efficient and fast optimization technique (PSO) is used to improve the allocation

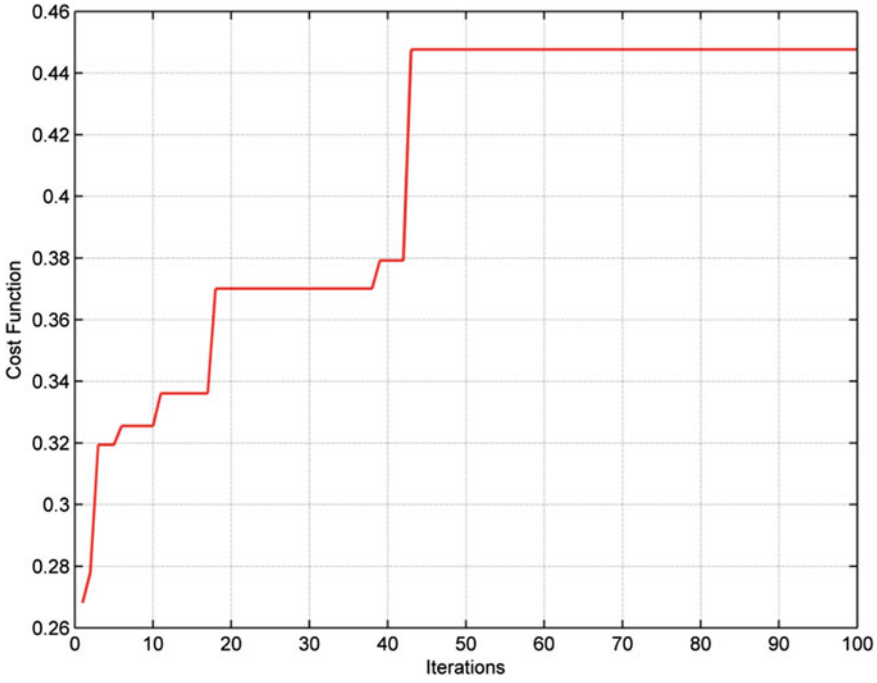


Fig. 10 PSO iterations for AMS volume (volume ratio) maximization

efficiency of GI by searching a solution which has greater AMS volume. Calculations of volume are performed using Bordignon [19] technique in Matlab software.

The ratio of the volumes between direct control allocation (DirCA) and GI control allocation is used as an objective function. The increase in the volume percentage will in fact improve the allocation efficiency of CA and given as:

$$\frac{V_{\psi}}{V_{\Phi}} \times 100$$

where V_{ψ} is the volume of GI solution and V_{Φ} is the volume of direct control solution, which is 100 % of the attainable moments. The ADMIRE aircraft's linear effectiveness matrix (D) is used to find a better GI solution. The aircraft's trim flight conditions are Mach = 0.3 and altitude h = 3000 m. Hence, the effective matrix is

$$v_d = B_e u$$

$$B_e = \begin{bmatrix} 0.7984 & -0.7984 & -4.5787 & -3.9413 & 3.9413 & 4.5787 & 2.6919 \\ 1.3841 & 1.3841 & -1.0906 & -1.7433 & -1.7433 & -1.0906 & 0.0046 \\ -0.3970 & 0.3970 & -0.2014 & -0.4256 & 0.4256 & 0.2014 & -1.6265 \end{bmatrix}$$

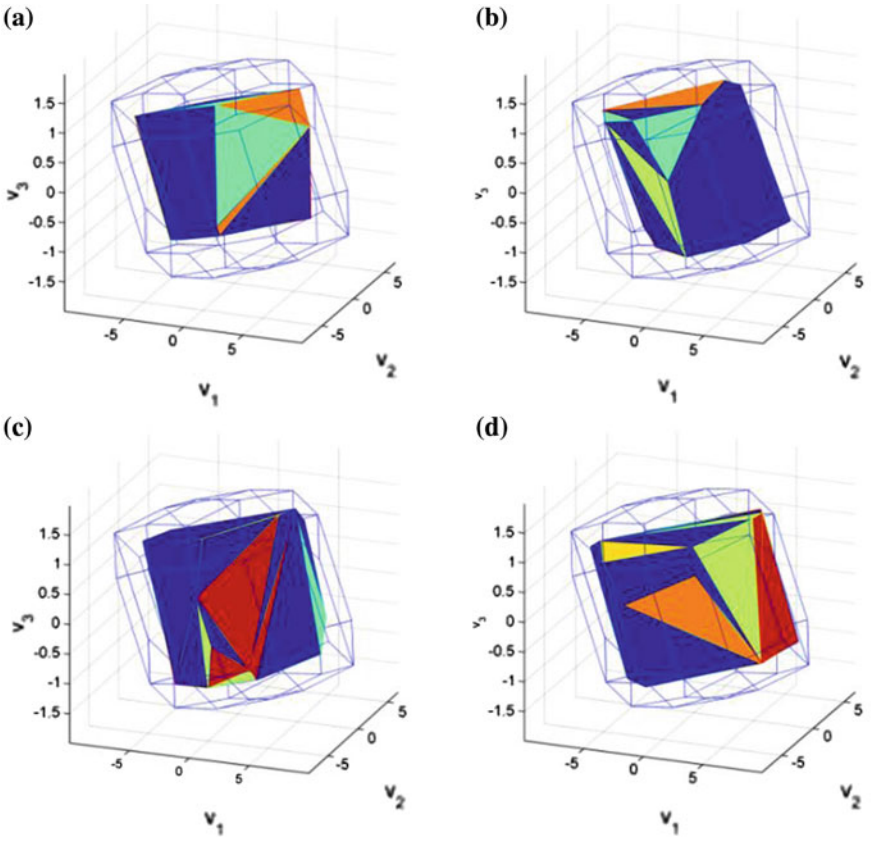


Fig. 11 AMS volumes at different iterations of PSO algorithm (**a**) 30 % (**b**) 36 % (**c**) 40 % (**d**) 45 %

Seven control surfaces i.e., right and left canards (δ_{rc}, δ_{lc}), right and left inner elevons (δ_{rie}, δ_{lie}), right and left outer elevons (δ_{roe}, δ_{loe}) and a rudder (δ_r) are used with the following given position limits

$$u_{\min} = [\ -55 \quad -55 \quad -25 \quad -25 \quad -25 \quad -25 \quad -30\ ],$$
$$u_{\max} = [\ \ 25 \quad \ \ 25 \quad \ \ 25 \quad \ \ 25 \quad \ \ 25 \quad \ \ 25 \quad \ \ 30\ ] \quad \text{(deg)}$$

The maximum volume of GI solution achieved through PSO algorithm is 45 % of the volume of AMS of Φ.

The Fig. 10 shows how fast PSO proceeded towards the optimized CA solution. The different volume changes at increasing iterations are shown in Fig. 11 and respective GI solutions are given in Table 3, where:

$$u = G_I\, v$$
$$I_{3\times 3} = B_e G_I$$
$$G_I = W_u (B_e W_u)^T \left(B_e W_u (B_e W_u)^T\right)^{-1}$$

Table 3 Optimized GI matrices for CA optimization at different volume percentages

GI solution at 30 %

$$G_I = \begin{bmatrix} 0.0334 & 0.2446 & -0.1494 \\ -0.0002 & 0.1217 & 0.1362 \\ -0.0476 & -0.0824 & -0.0811 \\ -0.0341 & -0.0911 & -0.0980 \\ 0.0334 & -0.0798 & 0.0732 \\ 0.0911 & -0.0966 & 0.1020 \\ 0.0266 & -0.0288 & -0.4776 \end{bmatrix}$$

GI solution at 36 %

$$G_I = \begin{bmatrix} 0.0095 & 0.2116 & -0.2773 \\ -0.0379 & 0.2252 & 0.2210 \\ -0.0428 & -0.0298 & -0.0499 \\ -0.0689 & -0.1033 & -0.1486 \\ 0.0350 & -0.0621 & 0.0817 \\ 0.0611 & -0.0684 & 0.0836 \\ 0.0285 & 0.0093 & -0.4164 \end{bmatrix}$$

GI solution at 40 %

$$G_I = \begin{bmatrix} 0.0337 & 0.1887 & -0.2327 \\ -0.0069 & 0.1771 & 0.2765 \\ -0.0557 & -0.0872 & -0.0701 \\ -0.0337 & -0.0832 & -0.0944 \\ 0.0443 & -0.1017 & 0.1268 \\ 0.0729 & -0.0700 & 0.0722 \\ 0.0264 & -0.0055 & -0.4150 \end{bmatrix}$$

GI solution at 45 %

$$G_I = \begin{bmatrix} 0.0194 & 0.2239 & -0.1620 \\ -0.0181 & 0.2667 & 0.1792 \\ -0.0550 & -0.0438 & -0.0793 \\ -0.0481 & -0.0656 & -0.0920 \\ 0.0488 & -0.0656 & 0.1009 \\ 0.0558 & -0.0419 & 0.0850 \\ 0.0299 & 0.0105 & -0.4607 \end{bmatrix}$$

12 LQ Controller Optimization

PSO based controller optimization is used very frequently in real control application because of its less computational intensive nature and fast convergence towards the solution [54–56]. In LQ control, both state and control penalty matrices (Q, R) are found systematically and quickly using PSO. Closed loop flight controller with control allocation is considered for response shaping. Integral time absolute error (ITAE) is used as objective function for determining the best pair of Q and R matrices.

PSO algorithm with inertia weight (w) and learning factors (c_1, c_2) explained in Sect. 9 is used here with ITAE cost function for evaluation of each particle to find the best possible feedback gain matrix. Following are the steps in closed loop flight controller optimization:

1. Select an initial swarm of particles.
2. For each particle solve Algebraic Riccati Equation (ARE) and calculate the feedback gain considering minimization of ITAE cost function for the desired command inputs.
3. If the current ITAE value is less than the previous best ITAE value then set the current value as the new best value (i best).
4. Now choose the particles with the minimum ITAE value of all the particles as the g best and calculate the particle velocity.
5. Using calculated particle velocity, find new particle position for Q and R matrices.
6. Stop, if either the maximum number of iteration is met or the optimal solution is found.

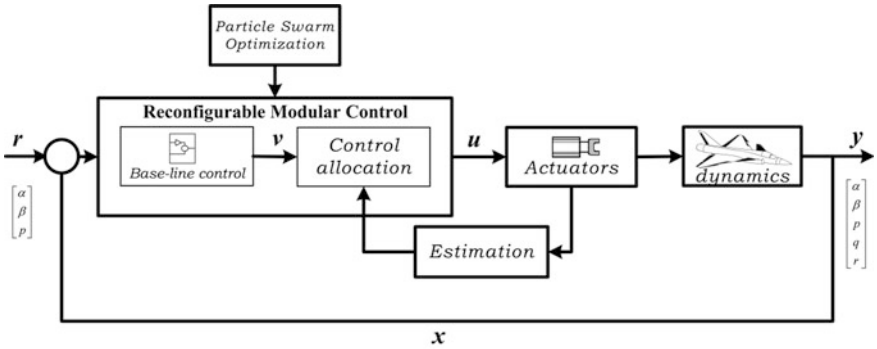


Fig. 12 Optimized modular feedback control system for ADMIRE using PSO

The improved LQ feedback controller performance is achieved by proper selection of Q and R weighting matrices having dimensions 5×5 and 3×3 respectively for ADMIRE aircraft. Recently, several intelligent optimization methods are investigated for determining the LQ weighting matrices, with closed-loop poles placement in complex left half plane [57]. The new poles placement achieved through PSO improves the stability index and minimizes the control effort. Thus employing fast intelligent optimization techniques for reconfigurable flight controller design is an innovative approach.

The Q and R are the positive diagonal weighting matrices of the following dimensions

$$Q = \begin{pmatrix} q_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & q_{55} \end{pmatrix}, \quad R = \begin{pmatrix} r_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & r_{33} \end{pmatrix}$$

13 Flight Control Example

ADMIRE aircraft simulation model is demonstrated here for improved simulation results with particle swarm optimization based reconfigurable controller shown in Fig. 12. ADMIRE linear model trimmed at, Mach 0.36 and altitude 3000 m is chosen to find the optimized gain matrices (Table 4).

The input and output configuration used in [37] is employed here for ADMIRE aircraft model, with the following state-space matrices of the system.

$$\dot{x} = A\,x + B_u\,u$$

Table 4 Open-loop system characteristics at Mach 0.36 and altitude 3000 m

States, $x(t)$	Eigenvalues	Damping (ζ)	Natural frequency (w_n) rad/sec
α	−3.04	1.00	3.04
β	1.38	−1.00	1.38
p	−1.45	1.00	1.45
q	−0.358 + 1.650i	0.212	1.69
r	−0.358 − 1.650i	0.212	1.69

where,

$$A = \begin{bmatrix} -0.8360 & 0.0079 & 0 & 0.9757 & 0 \\ 0 & -0.2019 & 0.0954 & 0 & -0.9853 \\ 0 & -17.4358 & -1.6017 & 0 & 0.5097 \\ 5.0110 & -0.0075 & 0 & -0.8230 & 0 \\ 0 & 1.2303 & -0.0889 & 0 & -0.3604 \end{bmatrix}$$

$$B_u = \begin{bmatrix} -0.0021 & -0.0021 & -0.0600 & -0.0972 & -0.0972 & -0.0600 & 0.0003 \\ -0.0061 & 0.0061 & 0.0035 & 0.0155 & -0.0155 & -0.0035 & 0.0470 \\ 0.9642 & -0.9642 & -6.6571 & -5.7541 & 5.7541 & 6.6571 & 3.8095 \\ 1.8939 & 1.8939 & -1.5950 & -2.5629 & -2.5629 & -1.5950 & 0.0067 \\ -0.5231 & 0.5231 & -0.2941 & -0.6388 & 0.6388 & 0.2941 & -2.3135 \end{bmatrix}$$

The open loop aircraft response at the given flight condition is unstable with a pole in right half s-plane (at 1.38). The insufficient damping of β, q, and r requires a feedback control system. For modular control design the actuator dynamics are neglected, and the introduction of virtual control concept as presented before is used to approximate the model as

$$B_u = B_v B_e$$

As the last three rows in B_u have dominant contribution of the control action on the system as compared to first two rows in B_u. So, the B_v and B_e are

$$B_v = \begin{bmatrix} 0_{2\times 3} \\ I_{3\times 3} \end{bmatrix}$$

$$B_e = \begin{bmatrix} 0.9642 & -0.9642 & -6.6571 & -5.7541 & 5.7541 & 6.6571 & 3.8095 \\ 1.8939 & 1.8939 & -1.5950 & -2.5629 & -2.5629 & -1.5950 & 0.0067 \\ -0.5231 & 0.5231 & -0.2941 & -0.6388 & 0.6388 & 0.2941 & -2.3135 \end{bmatrix}$$

Hence, the virtual control input, $v = B_e u$, contains the three angular accelerations (roll, pitch, and yaw) generated by the control surfaces. The weighting matrices Q = diag ([10 10 10 4 2]) and R = diag ([1.2 62.06 20.01]) are the

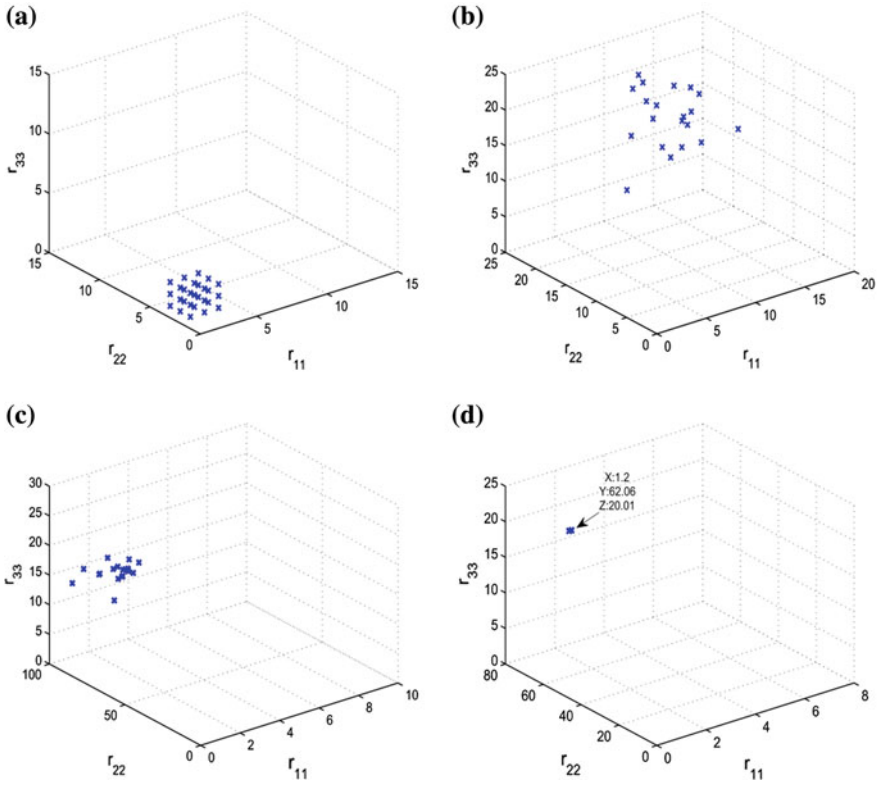


Fig. 13 Particle swarm searching for optimal R matrix (**a**) at start (**b**) after 20 iterations (**c**) after 40 iterations (**d**) at 100th iteration

optimized state and input matrices. The particle swarm search for the best R-matrix at different iterations is shown in Fig. 13. The PSO based optimized modular control response is shown in Fig. 14. The poles location and stability of closed-loop feedback control system is given in Table 5. All the poles are in LHP with reasonable damping ratio. The slowest time constant is 0.32 s.

Genetic algorithm (GA) is also used for reconfigurable modular flight controller optimization in ADMIRE aircraft simulation at this flight condition to explore the performance difference between PSO and GA. The optimization of control allocation to find the best GI solution results the same but the solution speed is relatively quite faster than the GA. Also the tuning of feedback modular controller is comparatively quite easy and takes half time as compare to GA based controller tuning.

The GA optimized and PSO based modular control response at nominal flight conditions is presented in Fig. 14. It can be easily seen that the particle swarm optimized controller has somewhat better response as compared to GA optimized

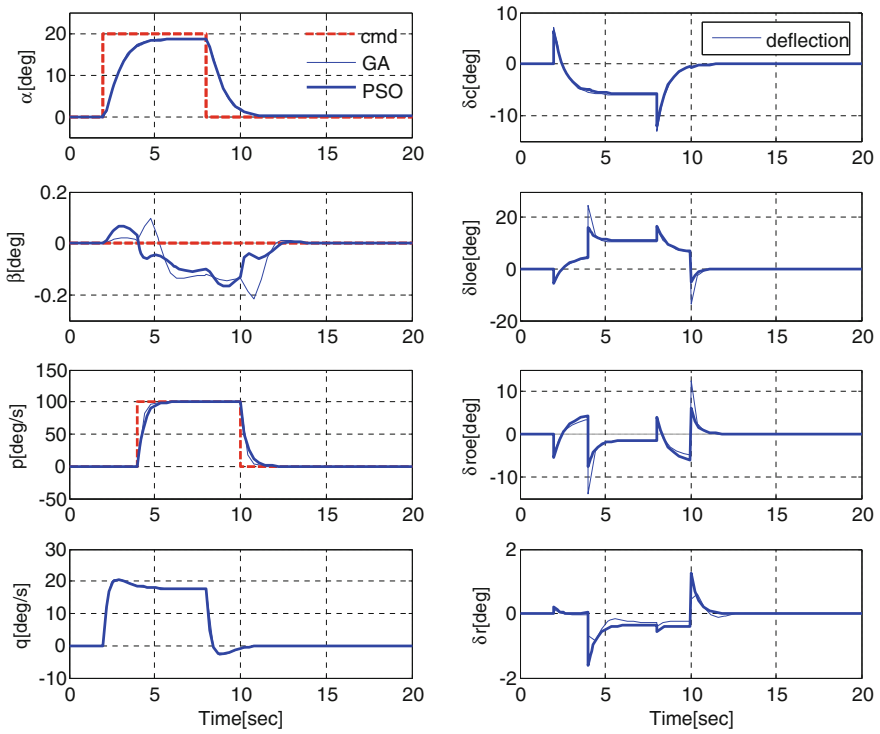


Fig. 14 Aircraft normal flight: reference trajectories (*dashed*) and actual trajectories for both GA optimized (*thin lines*) and PSO (*thick lines*) optimized controllers

Table 5 Modular feedback control system characteristics as optimized by PSO

States, $x(t)$	Eigenvalues	Damping (ζ)	Natural frequency (w_n) rad/sec
α	-1.39	1	1.39
β	$-1.08 + 1.8i$	0.514	2.10
p	$-1.08 - 1.8i$	0.514	2.10
q	-3.05	1	3.05
r	-3.08	1	3.08

controller. There is relatively less actuator deflections for the similar response achieved through GA optimized controller. Zero slideslip command tracking in PSO control system is improved in contrast to GA control system. The reconfigurable flight control performance is carried out by the redistribution of control effort using CGI CA, when post fault/failure information is available for redundant control surfaces is shown in Figs. 15 and 16.

It is evident that the two different fault cases in either canards or elevons cause an overshoot in pitch variables, angle of attack (α) and pitch rate (q). Because of

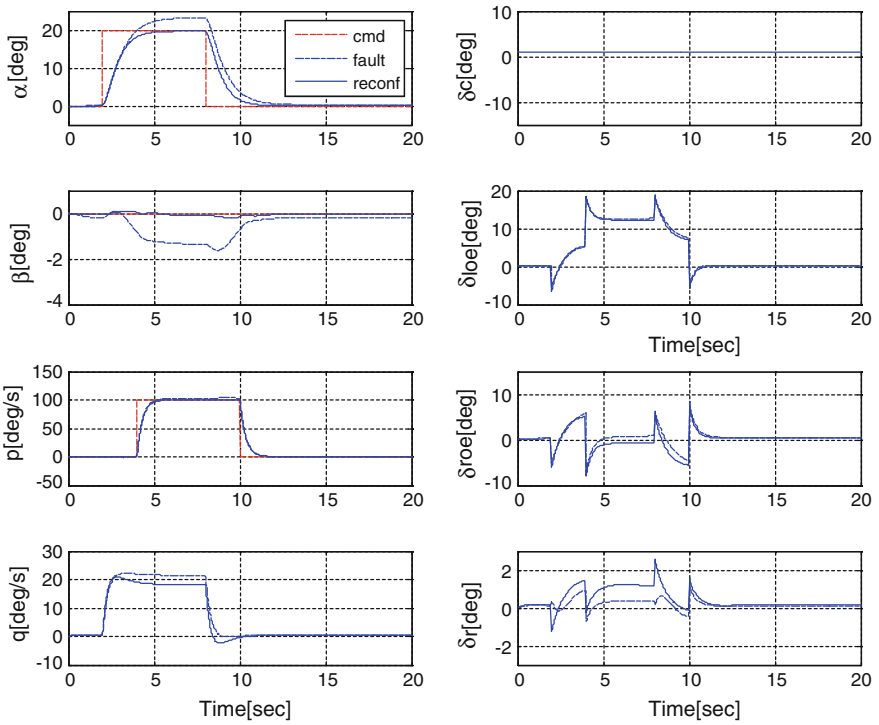


Fig. 15 Jammed canard fault: States and actuator deflections (fault at time = 0)

the available actuators redundancy in ADMIRE aircraft, pitch moment can be controlled by either the canard or elevons (left and right). In case of fault or failure, reconfigurable control action is performed through CA algorithm and healthy elevons can replace the damaged canard by re-distribution of control effort to elevons in achieving the desired pitch moment as shown in Figure). Left elevons saturation can be compensated by re-distributing the lost control effect to the right elevons and canard through CA as can be concluded from Figure.

14 Fast Embedded Control Allocation

In real flight control applications especially in high speed fighter aircrafts and rockets/missiles fast control allocation algorithm is the foremost requirement. Generally, in these fast moving vehicles, digital controller sampling frequency is 50–100 Hz which corresponds to a sampling time of 10–20 ms and CA is the part of complete control strategy. In safety-critical real-time embedded systems, optimization problem is solved many times before an optimal solution is available for

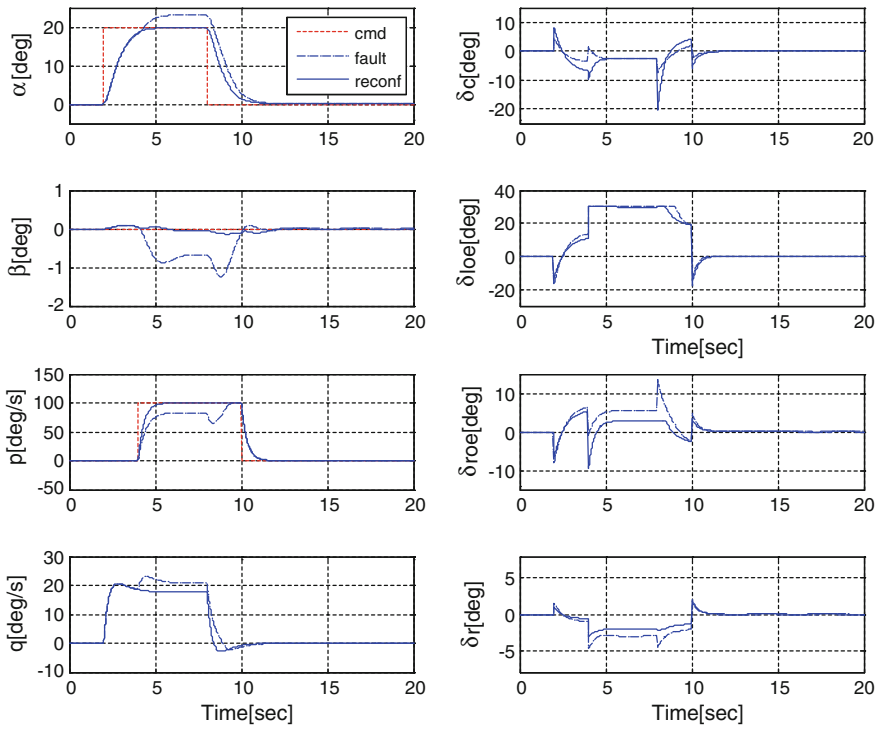


Fig. 16 Left elevons saturation fault: States and actuator deflections (fault at time = 2 s)

real implementation. It means that CA algorithm should be fast enough, so that it will provide optimal solution in few milliseconds or hundred of microseconds for real-time embedded application.

Optimization based CA strategies are now becoming a standard for efficient solution. Direct control allocation technique whose allocation efficiency is 100 % and based on the geometry of AMS, is also generally formulated to linear programming (LP) or quadratic programming (QP) problem for optimal and fast results [36, 58]. Generally, due to the design freedom available in optimization strategies a mixed optimization objective is recommended as.

$$\begin{aligned} &\min_{u} \left(\|W_u(u - u_d)\|_p^p + \gamma \|W_v (Bu - v)\|_p^p \right) \\ &\textit{subject to} \qquad u_{\min} \leq u \leq u_{\max} \end{aligned} \tag{8}$$

where W_u is the real input (u) weighting matrix, u_d is the desired real input, W_v is the virtual input (v) weighting matrix and p is the norm number. In Eq. 8, the first objective is the *control minimization* objective and second objective is the *error minimization* objective. The priority between the two objectives is defined through γ. Typically, the $\gamma \gg 1$ is high enough in above expression to prioritize the error

minimization objective. Exchange of these two objective terms in mixed optimization problem can also be used with $\gamma \ll 1$. Here, a convex optimization based real-time control allocation (RTCA) strategy is presented using disciplined convex programming techniques for mixed-objectives.

15 Disciplined Convex Programming

A new modeling methodology for formulating, analyzing, and solving convex optimization problems is introduced by Grant and Stanford research team [59] called as disciplined convex optimization. As the term "disciplined" suggests, this approach has set of conventions based on principles of convex analysis and inspired by practices that must be followed for convex programming. It is a set of library functions and sets whose convexity properties are already proved. Using these functions and sets for convex optimization become quite easy because most of the transformations necessary to convert disciplined convex program (DCP) description into solvable form can become fully automated.

The two main components of DCP are the *atom library*, containing functions and sets, or atoms, whose graphical properties i.e., shape/curvature (convex/concave/affine), range, and monotonicity is already declared and *ruleset* for atoms, variables, parameters, and numeric values based on convex analysis to produce convex results. Mathematical program developed in disciplined convex programming according to *ruleset* using elements of *atom library* can become conceptual and easy to understand. The details of *atom library* functions and *ruleset* for DCP can be found in Grant dissertation [60].

The simplified representation of convex optimization problem through DCP is suitable to understand the solution accuracy which can be achieved through convex optimization. DCP automatically transform natural representation of convex program to interior-point method for solution and then transform it back for results representation and understanding. Interior-point methods are quite fast for large problems but comparatively slower than active-set methods for small problems and medium size problems. A brief description of disciplined convex programming toolbox for Matlab is given in next section which is used here as development of real-time convex optimization based control allocation.

16 Implementation in cvx

The toolbox developed by Grant et al. [61] for DCP modeling framework in Matlab is called as cvx. cvx can be used to solve linear programs (LPs), quadratic programs (QPs), second-order cone programs (SOCPs), geometric programs (GPs), and semi-definite programs (SDPs) which all are subclasses of convex programming. In cvx tool some specific programming modes are also available

i.e., *SDP mode,* linear matrix inequalities (LMIs) and SDPs are expressed in more natural form for effective results, and *GP mode,* special functions and rules for geometric programming, including monomials, posynomials and generalized posynomials are used for efficient solution using convex form.

cvx is very effective learning and modeling environment for complex convex programming problems like l_p-norms specially for control allocation problem. The formulation of mixed optimization control allocation problem become greatly simple because of the presence of atomic library functions for norm minimization. Currently, cvx toolbox is using SeDuMi convex optimization solver. For installation of cvx toolbox refer to cvx user guide [61]. The mixed optimization based CA problem described by Eq. 8 can easily be implemented and verified for accuracy of solution for l_2-norm using following Matlab code:

```
cvx_begin
    variable u(n) % declare vector variable
minimize( norm(Wu*(u-ud),2) + gama*norm(Wv*(B*u-v),2) )
subject to u >= umin; u <= umax
cvx_end
```

In cvx specification *Wu*, *ud*, *gama*, *Wv*, *B*, *v*, *umin*, and *umax* are the inputs and should be provided before using above code for finding optimal distribution of virtual command on the redundant actuators.

17 Real-Time Control Allocation Using DCP

The cvx DCP modeling tool is not for real-time convex optimization due to its workflow. It is generally used for understanding and knowing the precision of the solution obtained through convex programming. It also requires extensive libraries and commercial software (i.e., Matlab) to run, which makes it inapplicable for real-time embedded systems. In our CA optimization problem, it provides solution in hundreds of millisecond which is not required and practically unusable for aircraft control.

Recently, a tool known as CVXGEN is developed by Mattingley et al. [62] for real-time code generation of convex optimization problems. It takes a high-level description of convex optimization problem and automatically generates fast, library-free C code that can be compiled and run for high speed execution. Code generated from CVXGEN is very efficient, robust and can run fast in few milliseconds or microseconds as in embedded form. The CVXGEN solvers are readily

available to use in embedded applications where fast reliable and optimal solution is required.

We have used CVXGEN for real-time control allocation (RTCA). The speed and accuracy of the CA solution is quite impressive and makes our approach dominating as compare to currently available CA strategies. The brief description of other promising control allocation approaches for CA optimization problem is presented in next section to compare them with convex optimization approach. The fast real-time implementation of these methods is performed in C using Matlab as a platform for evaluation.

18 Current Real-Time Control Allocation Methods

Some of the other competitive control allocation strategies widely used in CA problems are pseudo-inverse (PI) techniques, weighted least-squares (WLS), fixed-point method (FXP), interior-point method (IPM), and linear programming based CA method. These algorithms are also implemented in C code and evaluated with convex programming based CA algorithm for performance evaluation.

Simulation results: Comparison of simulation data is presented in Table 6 averaged over 1000 runs using Matlab commands (tic,toc) and error values of $\|Bu(t) - v(t)\|$ for F-18 aircraft data [63].

The effectiveness matrix B and position constraints of F-18 aircraft at 10,000 ft altitude, Mach 0.23, and 30° angle of attack are

$$B = \begin{bmatrix} 253.8 & -253.8 & 247.3 & -247.3 & 19.22 & -250.0 & 250.0 & 4.5 \\ -3801 & -3801 & -349.5 & -349.5 & 0.1681 & 1125 & 1125 & 0 \\ -16.81 & 16.81 & -92.51 & 92.51 & -382.7 & 0 & 0 & -750.0 \end{bmatrix} \times 1e^{-4}$$

$$u_{\min} = [-24 \quad -24 \quad -25 \quad -25 \quad -30 \quad -30 \quad -30 \quad -30]^T \times \frac{\pi}{180} \text{ rad}$$

$$u_{\min} = [10.5 \quad 10.5 \quad 42 \quad 42 \quad 30 \quad 30 \quad 30 \quad 30]^T \times \frac{\pi}{180} \text{ rad}$$

There are eight real control inputs u consist of left and right horizontal tails (u_1, u_2), left and right ailerons (u_3, u_4), ganged rudders (u_5), left and right thrust vectoring nozzles (u_6, u_7), working similarly as horizontal tails, and yaw thrust vectoring (u_8). The virtual control input v consists of three commanded moment coefficients C_l, C_m, and C_n each for roll, pitch, and yaw respectively. The virtual input profiles are shown in Figs. 17 and 18, which consist of 85 samples and the helical path radius of oscillating commands is 0.06, see [63].

After extensive evaluation of all real-time CA strategies, it can easily be concluded that convex programming based RTCA performed well in all aspects as can be seen in Table 6. Better load balancing can be seen in CVX_{RT} $(\text{L}_{\text{inf}} - \text{L}_1)$, where

Table 6 Control allocation results

Algorithm	Mean time (ms)	Max time (ms)	Mean error	Max error	Average control (°)	Mean iterations
WLS_NRT	0.28	0.45	4.19e-3	1.92e-2	18.11	3.82
WLS_RT	0.15	0.21	4.19e-3	1.92e-2	18.11	3.82
FXP_NRT	0.26	0.36	1.61e-2	2.41e-2	12.02	100
FXP_RT	0.13	0.16	1.61e-2	2.41e-2	12.02	100
IP_NRT	0.45	0.61	2.51e-2	3.78e-2	12.52	2.34
IP_RT	0.08	0.11	2.51e-2	3.78e-2	12.52	2.34
CGI_NRT	0.67	1.05	6.81e-3	5.01e-2	20.78	3.17
CGI_RT	0.17	0.24	6.81e-3	5.01e-2	20.78	3.17
CVX_{RT} (L_1-L_1)	0.11	0.15	4.79e-11	5.64e-10	15.31	6.88
CVX_{RT} (L_{inf}-L_1)	0.13	0.17	2.59e-11	3.54e-10	20.49	7.78

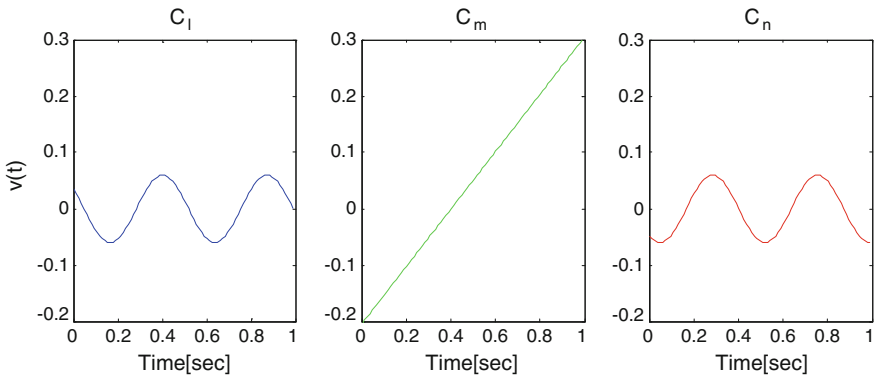


Fig. 17 Separate reference virtual control commands for CA algorithms

infinity norm is use for control minimization and L_1-norm for error minimization. Also the sensitivity analysis by Frost [64] is verified here.

Real-time implementation improves the timing performance to at-least twice with same error and average control properties. In comparison to solution time for real-time CA strategies CVX_{RT} solutions are the first choice and for error minimization, we can consider WLS_RT as a second option. If our criterion is simplicity of design then CGI_RT or FXP_RT can be a good choice.

19 Real-Time Flight Control Simulation

A complete reconfigurable flight controller is designed here considering military Level 1 flying qualities criteria (MIL-STD-8785C). Flight control system is based on proposed optimized reliable LQ control and optimization based real time

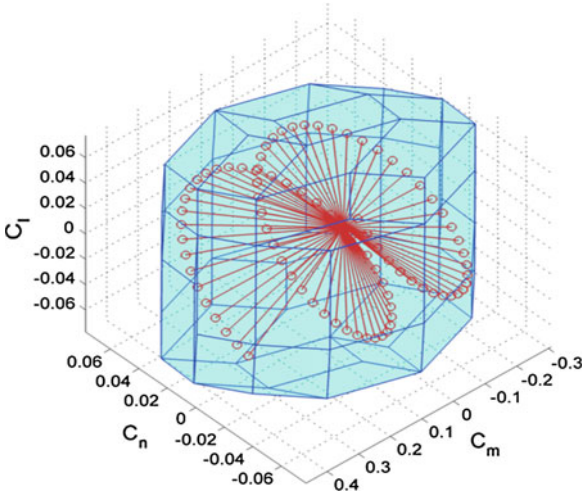


Fig. 18 Reference virtual command inputs in 3D AMS

control allocation. Also, the control system is implemented in hard real-time simulation development environment (xPC Target, MathWorks) for verification and validation (V&V).

20 Modular Flight Control System for ADMIRE

The proposed modular flight control system is shown in Fig. 19. The baseline controller is an integral LQ regulator which produces virtual command (v) consist of aerodynamic moment in roll, pitch, and yaw. This virtual command is then used as an input for control allocator which is based on optimization based RTCA strategies proposed and evaluated previously. The output of CA module is the optimized distribution of the v between the healthy control surfaces in the form of position deflections (u) considering the actuator limits. All the three flight conditions demonstrated before for ADMIRE aircraft are presented here for nominal aircraft trimmed at Mach 0.3 and Altitude 3000 m.

The robustness of the control strategy for redundant system and fault tolerance characteristics in case of actuator faults are demonstrated here considering the computational complexity in hard real-time environment.

The similar maneuvering flight control objectives discussed and demonstrated before are used here considering the flying qualities requirements of level 1. ADMIRE is a class IV aircraft with nonlinear dynamics. Following three reference commands are the control objectives

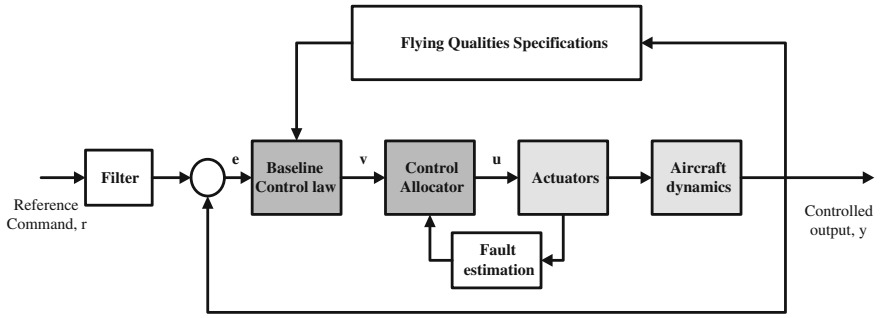


Fig. 19 Reconfigurable modular control configuration

$$\begin{aligned} \alpha &= \alpha_{ref} \\ \beta &= \beta_{ref} \\ p &= p_{ref} \end{aligned} \tag{9}$$

Because of the coupled dynamics of ADMIRE aircraft special attention is required for roll and yaw coupling (which should be minimized).

Following flying qualities characteristics are achieved by proper placement of feedback system poles using Q and R gain matrices of the ILQ regulator.

- In longitudinal short period mode the achieved damping ratio and frequency is $\zeta_{sp} = 0.66$ and $\omega_{nsp} = 3.5$ rad/sec
- In lateral-direction system, roll mode and dutch-roll mode specifications are met as $\tau_r = 0.4$ s, $\zeta_d = 0.83$, and $\omega_{nd} = 3.27$ rad/sec

The closed loop flight system performance can be seen in this chapter. Any control allocation strategy implemented in real time C code can be employed here for optimal distribution of virtual command among the redundant control surfaces considering their respective position limits. The nominal conditions for mixed optimization CA objective (see Eq. 8) with $\gamma = 1e3$ and following gain matrices are used here.

$$W_v = diag\,([1\ 1\ 1])$$

$$W_u = diag([1\ 1\ 1\ 1\ 1\ 1\ 1])$$

By modifying the respective entries in W_v matrix, we can prioritize the respective control objective given in Eq. 8. The control reconfiguration is performed by redistribution of control effort through W_u in fault and failure case. Several different flight simulations are performed for control system properties evaluation in next section.

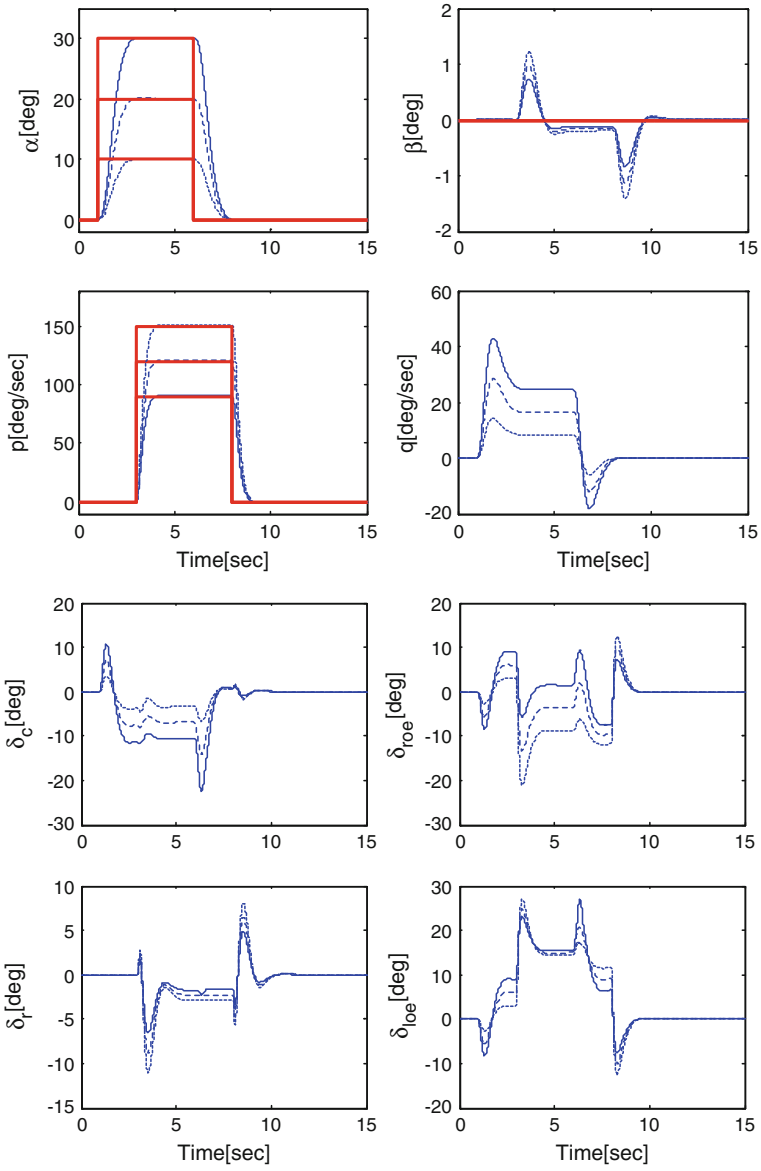


Fig. 20 Illustration of controller tracking: Reference commands (*thick red lines*) at $\alpha_{ref} = 10$, 20, and 30° and $p_{ref} = 150$, 120, and 90 °/sec with controller response in *dotted lines*, *dashed lines*, and *solid lines* respectively

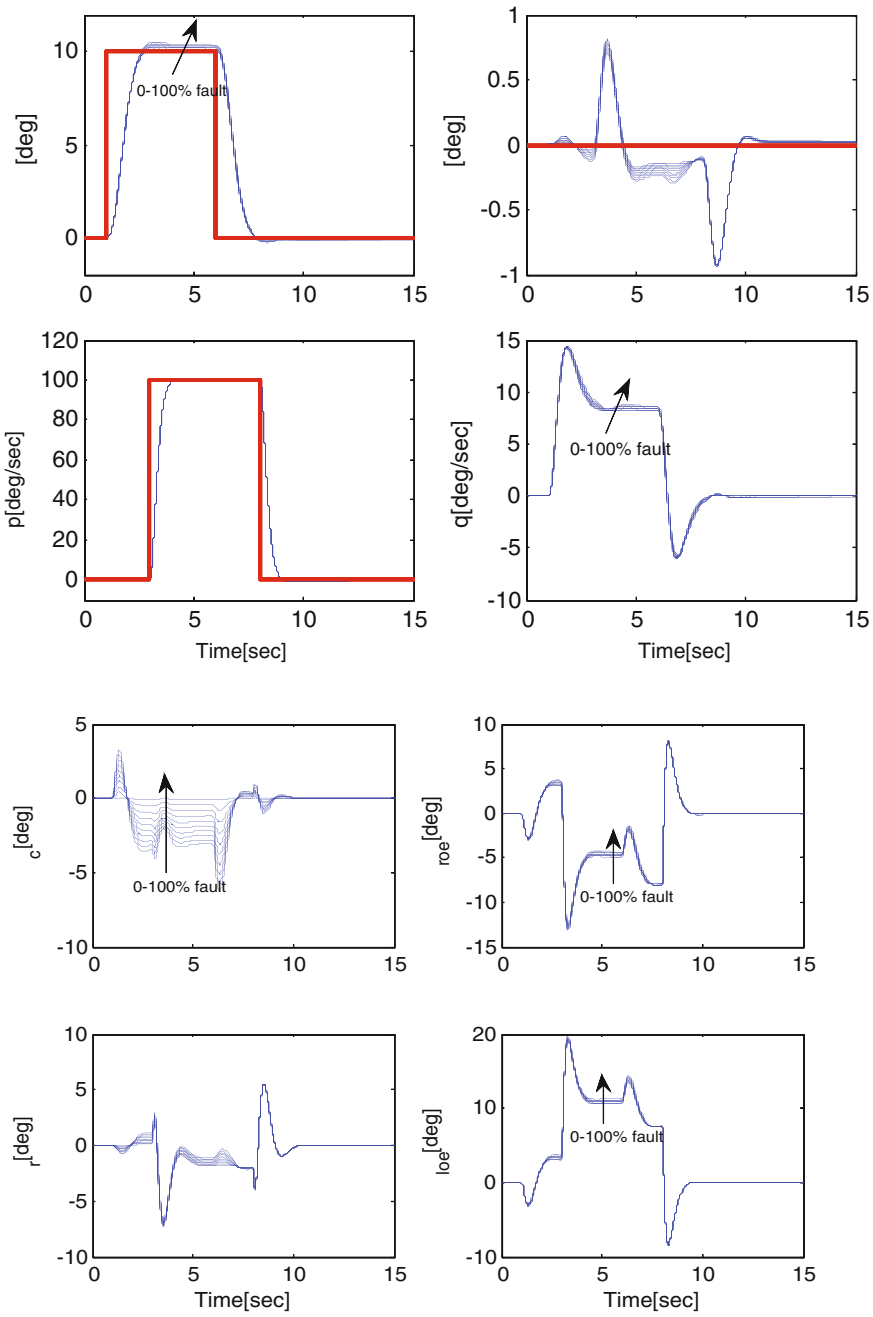


Fig. 21 Canard loss of effectiveness fault: Reference command (*thick red lines*). Flight simulation response (*thin solid lines*), where fault in canard increases from 0 to 100 % (*thin lines*)

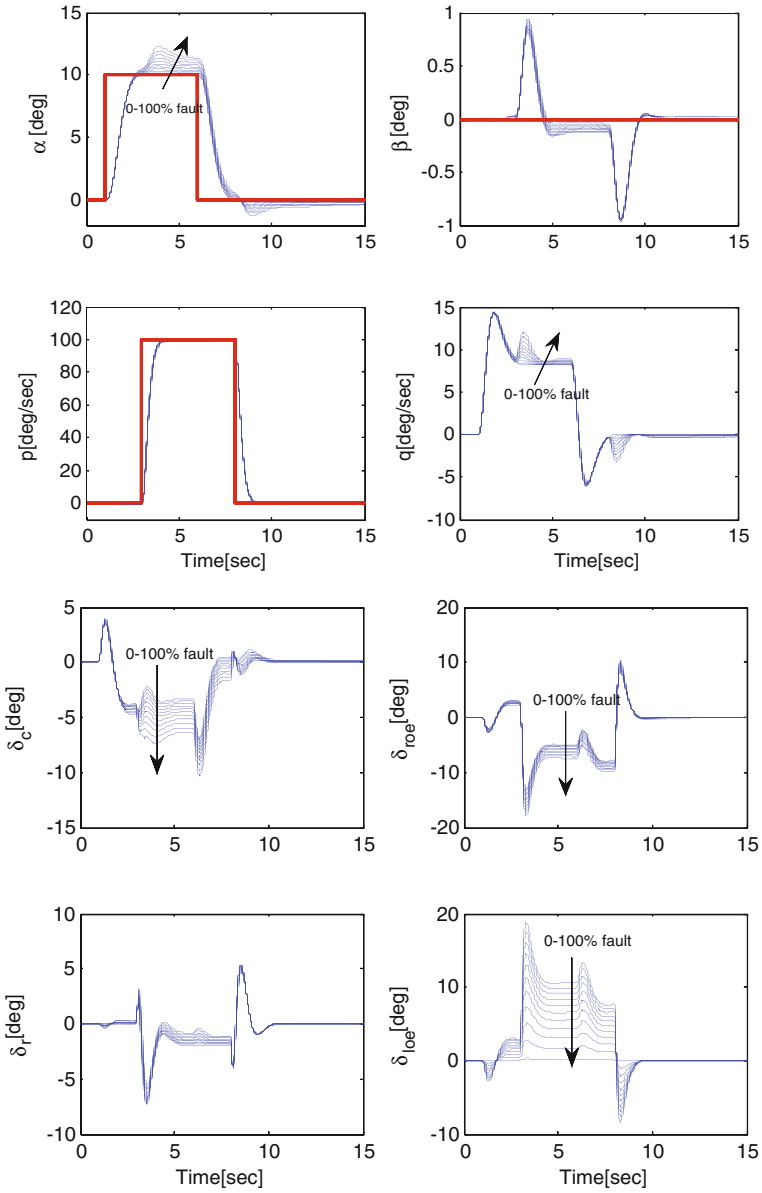


Fig. 22 Left elevon loss of effectiveness fault: Commanded trajectory (*thick red lines*). Flight simulation response (*thin solid lines*), where fault in left elevon increases from 0 to 100 % (*thin lines*) at the rate of 10 % increase

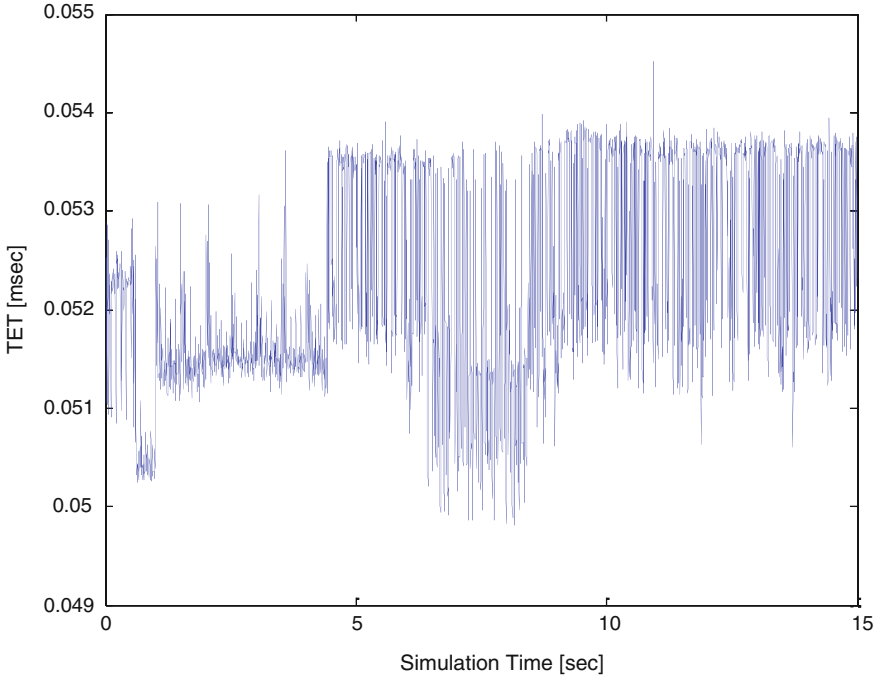


Fig. 23 Task execution time (TET) at each sample time for complete simulation: *Top* TET plot with first initialization sample. *Bottom* TET plot ignoring first sample for clear picture of timing requirements

21 Simulation Results

The following simulation scenarios are evaluated at different conditions.

a. ***Variations in reference commands:*** In Fig. 20 different angle-of attack reference commands are applied for different roll rate commands to evaluate the tracking performance of flight controller for higher angles of attack. With 10 deg increase in commanded α and 30 deg/sec decrease in respective commanded p all states are perfectly tracked within the actuators limits.
b. ***Canard fault:*** Various levels of canard faults (from 0 to 100 %) are introduce by varying the effectiveness in fault estimation matrix as explained before. Figure 21 shows that the over-shoot in α tracking due to loss of effectiveness fault in canard is avoided by using CA block through systematically redistributing the control demand between the left and right elevons.
c. ***Left elevon fault:*** The similar fault levels with increment of 10 % increase are included in left elevon control surface as discussed before. The noticeable degradation in Fig. 22 for α command tracking is due to the less available

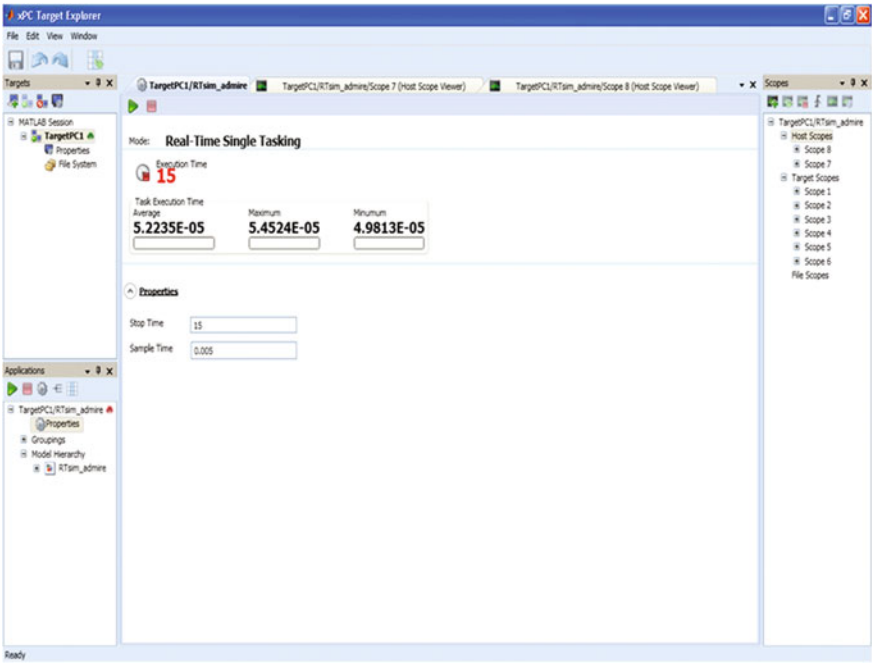


Fig. 24 Host PC graphical using interface (GUI) for data logging, parameter tuning, and controlling target PC simulation: xPC Target explorer screen capture after completion of simulation

redundancy for elevon control surface. Elevon control surface is used for both pitch and roll tracking which cannot be fully replaced by canard control surface which can only be used for pitch tracking.

22 Real Time Simulation Results

ADMIRE modular flight control simulation block diagram shown in Fig. 19 is implemented in real-time for evaluation of algorithm timing and response. The consistency and the issues between the Simulink model, Matlab code and C-code in our simulation is handled by the automated code generator. The generated code is compatible, readable, reliable and with no or least error rate as compared to manual code implementation.

Figure 23 shows the TET in milli-seconds versus the each sample time in 15 s simulation. The maximum TET is approximately 55 μs, which shows that we can easily execute our RT simulation at lower sample time but usually flight controllers are sampled at 10–20 ms.

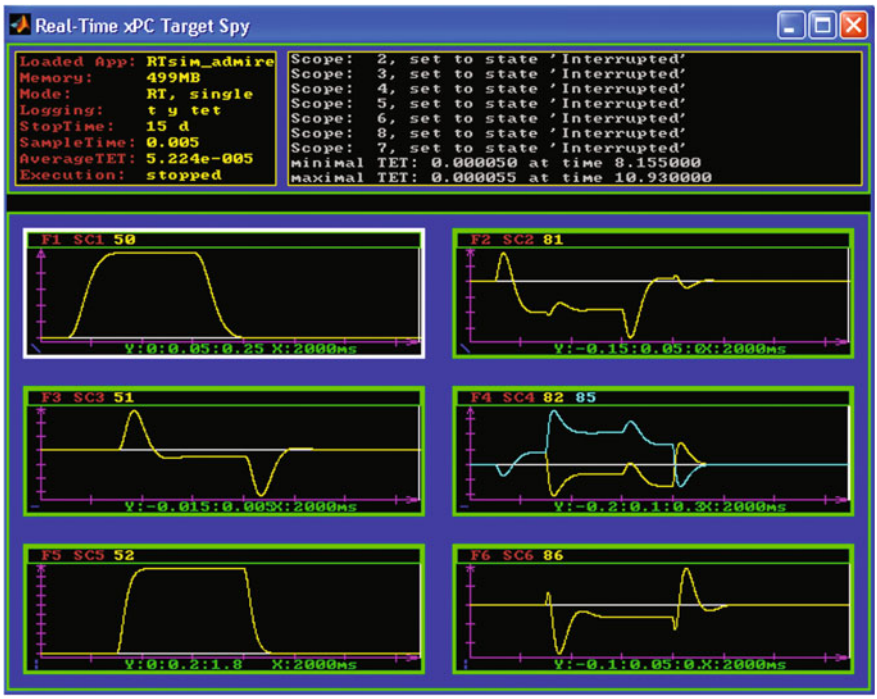


Fig. 25 Target PC screen capture after simulation stopped

In Fig. 24, the host PC interface is shown, where target simulation properties are shown with minimum, maximum, and average TET. The Matlab 8.0 is selected here for simulation development and experimentation in xPC Target environment. The top left portion shows the host-target connection status, whereas the bottom window shows the simulink application properties and model hierarchy. The right most window shows the available host and target scopes with their respective controls at the top.

The target PC flight status monitoring window at two different time instants are shown in Fig. 25. In scope 1, scope 3, and scope 5 (SC1,SC3,SC5) flight states α, β, and *p*are shown respectively. Whereas, in scope 2, scope 4 and scope 6 (SC2,SC4,SC6) aircraft control effectors canard, left elevon, right elevon and rudder position status is shown respectively.

The real-time flight simulator developed here can be used for controlled aircraft flight performance evaluation and HIL testing. Through the inclusion of environmental disturbance models, actual pilot interface, actual aircraft sensors and actuator load simulators, a real aircraft control performance characteristics can be determine precisely before deploying the embedded code in embedded flight control computer (FCC).

23 Conclusion

In this chapter, an optimized modular fault tolerant flight control design approach is discussed which is a mandatory part of an aerospace CPS. The modular configuration for flight control is composed of inner and outer feedback loops. The outer baseline controller feedback loop ensures the stability and robustness while inner reconfigurable design is responsible for the fault-tolerance against actuator faults/failures. Our results show that the proposed strategy guarantees augmented autonomy and intelligence on board aircraft for real time decision, reconfiguration and fault tolerant control. An special reference is made to real time control allocation (RTCA) by using PSO as well as genetic algorithms for an ADMIRE benchmark aircraft model. This approach is found to offer a fault tolerant approach for an over actuated aircraft in case of actuator faults. Moreover, for RT implementation of embedded optimization, cvx is used as a tool for convex optimization. Simulation results show the fault tolerant response comparison for varying reference tracking commands under canard and left elevon faults. Task execution time at each sample time is logged for complete simulation. A successful RT implementation is demonstrated by using xPC target explorer.

References

1. K. Butts, N. Sivashankar, J. Sun, Feedforward and feedback design for engine idle speed control using l1 optimization, in *Proceedings of the American Control Conference, 1995*, pp. 2587–2590
2. A. Tarau, B. De Schutter, H. Hellendoorn, Centralized versus decentralized route choice control in dcv-based baggage handling systems, in *International Conference on Networking, Sensing and Control, 2009. ICNSC'09*, 2009, pp. 334–339
3. W. Wang, D.E. Rivera, K.G. Kempf, Centralized model predictive control strategies for inventory management in semiconductor manufacturing supply chains, in *Proceedings of the American Control Conference, 2003*, pp. 585–590
4. C. Vermillion, J. Sun, K. Butts, Modeling, control design, and experimental validation of an overactuated thermal management system for engine dynamometer applications. IEEE Trans. Control Syst. Technol. **17**, 540–551 (2009)
5. M.S. Kim, C.S. Chung, The robust flight control of an uav using mimo qft: ga-based automatic loop-shaping method, in *Systems Modeling and Simulation: Theory and Applications*, pp. 467–477 (2005)
6. Härkegård, Ola, and S. Torkel Glad. Resolving Actuator Redundancy—optimal control vs. control allocation. Automatica **41**(1), 137–144 (2005)
7. O. Harkegard, Dynamic control allocation using constrained quadratic programming. J. Guid. Control Dyn. **27**, 1028–1034 (2004)
8. Y. Luo, A. Serrani, S. Yurkovich, D.B. Doman, M.W. Oppenheimer, Model predictive dynamic control allocation with actuator dynamics, in *Proceedings of the American Control Conference, 2004*, pp. 1695–1700
9. J. Tjonnas, T.A. Johansen, Optimizing adaptive control allocation with actuator dynamics, in *2007 46th IEEE Conference on Decision and Control, 2007*, pp. 3780–3785

10. T.A. Johansen, T.P. Fuglseth, P. Tøndel, T.I. Fossen, Optimal constrained control allocation in marine surface vessels with rudders. Control Eng. Pract. **16**, 457–464 (2008)
11. O. Sørdalen, Optimal thrust allocation for marine vessels. Control Eng. Pract. **5**, 177–185 (537)
12. S.P. Berge, T.I. Fossen, Robust control allocation of overactuated ships; experiments with a model ship, in *Proceedings of the 4th IFAC Conference on Manoeuvring and Control of Marine Craft*, vol. 537, pp. 20–25
13. T.A. Johansen, Optimizing nonlinear control allocation, in *43rd IEEE Conference on Decision and Control, 544, CDC. 544*, vol. 4, pp. 3435–3440
14. T.A. Johansen, T.I. Fossen, P. Tøndel, Efficient optimal constrained control allocation via multiparametric programming. J. Guidance Control Dyn. **28**, 506–515
15. K.P. Lindegaard, T.I. Fossen, Fuel-efficient rudder and propeller control allocation for marine craft: experiments with a model ship. IEEE Trans Control Syst. Technol. **11**, 850–862
16. R.H. Shertzer, D. Zimpfer, P.D. Brown, Control allocation for the next generation of entry vehicles, in *AIAA Guidance, Navigation, and Control Conference and Exhibit*
17. Y. Luo, A. Serrani, S. Yurkovich, M.W. Oppenheimer, D.B. Doman, Model-predictive dynamic control allocation scheme for reentry vehicles. J. Guidance Control Dyn. **30**, 100–113
18. O. Härkegård, Efficient active set algorithms for solving constrained least squares problems in aircraft control allocation, in *Proceedings of the 41st IEEE Conference on Decision and Control, 542*, vol. 2, pp. 1295–1300
19. K.A. Bordignon, Constrained control allocation for systems with redundant control effectors, Ph.D. Dissertation, Virginia Polytechnic Institute and State University, 1996
20. J.M. Buffington, Tailless aircraft control allocation. Tech. Rep., DTIC Document, 1997
21. J.H. Plumlee, D.M. Bevly, A.S. Hodel, Control of a ground vehicle using quadratic programming based control allocation techniques, in *Proceedings of the American Control Conference, 2004*, vol. 5, pp. 4704–4709
22. B. Schofield, T. Hagglund, A. Rantzer, Vehicle dynamics control and controller allocation for rollover prevention. IEEE International Conference on Control Applications Computer Aided Control System Design, 546, IEEE International Symposium on Intelligent Control, IEEE, pp. 3–8
23. J. Tjoennas, T.A. Johansen, Adaptive optimizing dynamic control allocation algorithm for yaw stabilization of an automotive vehicle using brakes, in *14th Mediterranean Conference on Control and Automation, 546. MED'06*, 546, pp. 1–6
24. P. Tøndel, Constrained optimal control via multi-parametric quadratic programming, Ph.D. thesis, Norwegian University of Science and Technology, 543
25. J. Tjonnas, T.A. Johansen, Stabilization of automotive vehicles using active steering and adaptive brake control allocation. IEEE Trans. Control Syst. Technol. **18**, 545–558 (2010)
26. D.B. Doman, M.W. Oppenheimer, D. Sigthorsson, Dynamics and control of a biomimetic vehicle using biased wingbeat forcing functions: Part ii: Controller. AIAA, Washington, **1024** (2010)
27. K.S. Jung, Y.S. Baek, Force distribution of a six-legged walking robot with high constant speed. J. Mech. Sci. Technol. **14**, 131–140 (2000)
28. S. Sreenivasan, K. Waldron, S. Mukherjee, Globally optimal force allocation in active mechanisms with four frictional contacts. J. Mech. Des. **118**(3), 353–359 (1996)
29. M.P.J. Fromherz, W.B. Jackson, Force allocation in a large-scale distributed active surface. IEEE Trans. Control Syst. Technol. **11**, 641–655 (2003)
30. K. Bordignon, J. Bessolo, Control allocation for the X-35b, in *2002 Biennial International Powered Lift Conference and Exhibit*, 2002
31. J.B. Davidson, F.J. Lallman, W.T. Bundick, Real-time adaptive control allocation applied to a high performance aircraft, in *5th SIAM Conference on Control & Its Application, 2001*, pp. 1211–1229
32. O. Harkegard, Backstepping and control allocation with applications to flight control, Ph.D. dissertation, Linkoping universitet, 2003

33. J.Q. Leedy, Real-Time Moment Rate Constrained Control Allocation for Aircraft with a Multiply-Redundant Control Suite, Ph.D. thesis, Virginia Polytechnic Institute and State University, 1998
34. W.C. Durham, Computationally efficient control allocation. J. Guid. Control Dyn. **24**(3), 519–524 (2001)
35. K.R. Scalera, A comparison of control allocation methods for the F-15 ACTIVE research aircraft utilizing real-time piloted simulations, Ph.D. thesis, Virginia Polytechnic Institute and State University, 1999
36. M. Bodson, Evaluation of optimization methods for control allocation. J. Guidance Control Dyn. **25** (2008)
37. A.H. Khan, Z. Weiguo, Z.H. Khan, S. Jingping, Evolutionary computing based modular control design for aircraft with redundant effectors. Procedia Eng. **29**, 110–117 (2012)
38. A.H. Khan, Z. Weiguo, S. Jingping, Z.H. Khan, Optimized reconfigurable modular flight control design using swarm intelligence. Procedia Eng. **24**, 621–628 (2011)
39. M.A. Bolender, D.B. Doman, M.W. Oppenheimer, Application of piecewise linear control allocation to reusable launch vehicle guidance and control, in *14th Mediterranean Conference on Control and Automation, 2006. MED'06*, 2006, pp. 1–10
40. J.A.M. Petersen, Algorithms for optimal real-time control allocation, Ph.D. dissertation, 2003
41. J.A.M. Petersen, M. Bodson, Interior-point algorithms for control allocation. J. Guid. Control Dyn. **28**, 471–480 (2005)
42. Y. Hattori, K. Koibuchi, T. Yokoyama, Force and moment control with nonlinear optimum distribution for vehicle dynamics, in *Proceedings of the 6th International Symposium on Advanced Vehicle Control* (2002), pp. 595–600
43. R.C. Eberhart, Y. Shi, J. Kennedy, Swarm intelligence, Morgan Kaufmann, 2001
44. J. Kennedy, R. Eberhart, Particle swarm optimization, in *Proceedings of the IEEE International Conference on Neural Networks*, vol. 4 (1995), pp. 1942–1948
45. Z.L. Gaing, A particle swarm optimization approach for optimum design of pid controller in avr system. IEEE Trans. Energy Convers. **19**, 384–391 (2004)
46. N. Jin, Y. Rahmat-Samii, Parallel particle swarm optimization and finite-difference timedomain (pso/fdtd) algorithm for multiband and wide-band patch antenna designs. IEEE Trans. Antennas Propag. **53**, 3459–3468 (2005)
47. Y. del Valle, G.K. Venayagamoorthy, S. Mohagheghi, J.C. Hernandez, R.G. Harley, Particle swarm optimization: basic concepts, variants and applications in power systems. IEEE Trans. Evol. Comput. **12**, 171–195 (2008)
48. R. Poli, Analysis of the publications on the applications of particle swarm optimization. J. Artif. Evol. Appl. **2008**, 3 (2008)
49. Y. Shi, R. Eberhart, A modified particle swarm optimizer, in *Proceedings of the 1998 IEEE International Conference on Evolutionary Computation* (IEEE World Congress on Computational Intelligence, 1998), pp. 69–73
50. R.C. Eberhart, Y. Shi, Tracking and optimizing dynamic systems with particle swarms, in *Proceedings of the 2001 Congress on Evolutionary Computation*, vol. 1 (2001), pp. 94–100
51. P. Fourie, A. Groenwold, The particle swarm optimization algorithm in size and shape optimization. Struct. Multidiscip. Optim. **23**, 259–267 (2002)
52. Y.L. Zheng, L.H. Ma, L.Y. Zhang, J.X. Qian, On the convergence analysis and parameter selection in particle swarm optimization, in *Proceedings of the 2003 International Conference on Machine Learning and Cybernetics*, vol. 3 (2003), pp. 1802–1807
53. J. Kennedy, R. Eberhart, Particle swarm optimization, in *Proceedings of the IEEE International Conference on Neural Networks*, vol. 4 (1995), pp. 1942–1948
54. M. Nasri, H. Nezamabadi-Pour, M. Maghfoori, A PSO-based optimum design of pid controller for a linear brushless dc motor. World Acad. Sci., Eng. Technol. **26**, 211–215 (2007)
55. C.N. Ko, C.J. Wu, A PSO-tuning method for design of fuzzy PID controllers. J. Vib. Control **14**, 375–396 (2008)

56. M.I. Solihin, W. Akmeliawati, R. Akmeliawati, PSO-based optimization of state feedback tracking controller for a flexible link manipulator, in Proceedings of the International Conference of Soft Computing and Pattern Recognition, 2009 (SOCPAR'09) (2009), pp. 72–76
57. S.A. Ghoreishi, M.A. Nekoui, S.O. Basiri, Optimal design of LQR weighting matrices based on intelligent optimization methods. IJIIP: Int. J. Intell. Inf. Process., AICIT **2**, 63–74 (2011)
58. D. Enns, Control allocation approaches, in *Proceeding of the AIAA Guidance, Navigation, and Control Conference* (1998)
59. M. Grant, S. Boyd, Y. Ye, Disciplined convex programming, in *Global Optimization* (Springer, New York, 2006), pp. 155–210
60. M. Grant, Disciplined convex programming. Ph.d. dissertation, Stanford University (2004)
61. M. Grant, S. Boyd, Y. Ye, cvx users' guide. Technical Report Build 711, Citeseer (2009), http://citeseerx.ist.psu.edu/viewdoc/download
62. J. Mattingley, S. Boyd, CVXGEN: a code generator for embedded convex optimization, in *Optimization and Engineering* (2012), pp. 1–27
63. W.C. Durham, K.A. Bordignon, Multiple control effector rate limiting. J. Guid. Control Dyn. **19**, 30–37 (1996)
64. S.A. Frost, M. Bodson, Resource balancing control allocation. Am. Control Conf. (ACC) **2010**, 1326–1331 (2010)

TS Fuzzy Approach for Fault Detection in Nonlinear Cyber Physical Systems

Muhammad Abid, Abdul Qayyum Khan, Muhammad Rehan and Haroon-ur-Rasheed

A fault is deviation of a characteristic property of a system from acceptable, standard condition. That is, some components of the process under consideration may not respond according to their desired behavior, for example, there may be blockage in pipes, the valves controlling the flow rates of fluids may stuck up, the gears in mechanical systems may be broken, the sensors measuring position, temperature, pressure or any other quantity may give incorrect results. Faults may have several effects on plants, for example, if there is a sensor fault, the plant will be operating away from its nominal/optimal operating point resulting into reduced efficiency of the plant and lower quality of product. Furthermore, in dynamic systems, if the sensor measurements are also being fed back, the stability of the process may be at risk. Likewise, if there is fault in actuator of the process, it will stop responding (or will respond incorrectly) to the inputs; this will reduce the availability of the plant. All these factors cause economic losses. In certain safety critical processes, faults not only cause economic losses, but may also be fatal. For example, there was a fatal accident in the isomerization unit of British Petroleum operated oil refinery in Texas, USA. Five people were killed and over 170 harmed as a result of the explosion [30]. The investigations showed that the reason was a

M. Abid (✉) · A. Q. Khan · M. Rehan · Haroon-ur-Rasheed
Department of Electrical Engineering, Pakistan Institute of Engineering and Applied Sciences, Nilore, Islamabad, Pakistan
e-mail: mabid@pieas.edu.pk

A. Q. Khan
e-mail: aqkhan@pieas.edu.pk

M. Rehan
e-mail: rehan@pieas.edu.pk

Haroon-ur-Rasheed
e-mail: haroon@pieas.edu.pk

Z. H. Khan et al. (eds.), *Computational Intelligence for Decision Support in Cyber-Physical Systems*, Studies in Computational Intelligence 540,
DOI: 10.1007/978-981-4585-36-1_14,

faulty sensor in the unit which was stuck to measure a constant liquid level despite it was continuously increasing. The disastrous effects of faults may be avoided/minimized if faults are quickly detected and remedial actions are timely taken. For example, there was an accident in American Airline DC10 which crashed at Chicago O'Hare International Airport. Later studies indicated that the accident could have been avoided [37]. Boeing 747-200F lost both of its engines while taking off from Schiphol International Airport. The post-accident analysis showed that this accident could have been avoided if the faults were detected and controllers were reconfigured accordingly [27].

The study of fault detection techniques emerged as a research area in 1970s with the arrival of modern computers. The research was initiated by control community and observer based approaches were utilized. Later on, researchers from different engineering fields tackled the problem and statistical techniques, signal processing techniques, artificial intelligence including neural networks, swarm networks, genetic algorithms were utilized for fault detection. In addition to that, the combinations of different approaches were proposed. Among these interdisciplinary areas, Takagi-Sugeno Fuzzy models have received considerable interest for fault detection in nonlinear dynamic systems [6]. A comprehensive study of TS fuzzy approach for fault detection can be found in [38]. TS Fuzzy approach has been successfully applied for fault detection to many practical systems. Here we name only a few applications, [32] applied TS Fuzzy models for fault detection in fuel cells, and the sensor faults are successfully detected. TS fuzzy approach has been applied to fault detection of many electrical systems, among them, [29] presented fault detection of transformers, and [26] presented the application of TS fuzzy approach for fault detection of induction motors. The problem of fault-tolerant tracking control for near-space-vehicle attitude dynamics using TS fuzzy approach is discussed in [22]. Application of fuzzy approach for fault detection in gas turbines using neuro-fuzzy approach is handled in [33].

Cyber physical systems (CPS) are physical systems with integrated computational and communicational resources. Some examples of CPS are power generation and distribution networks, unmanned aerial vehicles and SCADA systems. Fault detection in cyber physical systems has been discussed in a few research articles [34–36, 40, 45].

This chapter will present some basic concepts in fault detection and fuzzy logic. Observer-based approaches for fault detection in linear dynamical systems will be introduced. The motivation of using the TS fuzzy approach will be presented by briefly describing the fault detection techniques for nonlinear dynamic systems and their shortcomings. Some recent advances in TS fuzzy approach for fault detection will be presented. Application of TS fuzzy approach to fault detection in Cyber Physical systems will be described. Some future directions for possible extension of the existing TS Fuzzy approaches will be highlighted.

1 Basic Concepts in Fault Diagnosis

In this section, some basic concepts in fault diagnosis will be introduced.

1.1 Plant

Plant or System is defined as the process to be monitored. For the purpose of control or fault detection, large systems are generally divided into small subsystems and fault detection/control algorithms are applied to each subsystem. In the sequel, the terms plant, system or process will be used interchangeably. Any system is generally composed of plant components, actuators and sensors, as shown in Fig. 1. Plant components are, for example, tanks and pipes in chemical industry, wings of aircraft to be controlled, and the control rods in a nuclear reactor. Actuators are used to convert control signals in large actuating inputs, for example, motors, pumps, etc. Sensors are used to measure the quantity to be controlled, for example, thermocouples, pressure sensors, optical encoders, and so on. Plants may be either static or dynamic in nature. Generally, fault detection in dynamic systems is difficult compared to fault detection in static systems.

1.2 Fault

Fault is defined as an un-permitted deviation of at least one characteristic property of a system from standard condition. An analogous term is failure which is used to indicate complete disruption of the system. However, as far as detection is concerned, the same techniques are used to detect faults and failures, therefore, the two terms are sometimes used interchangeably.

Faults may be classified according to their source, that is, faults appearing in plant components are called component faults, those appearing in the sensors are called sensor faults and those in actuators are called actuator faults. Figures 2 and 3 show some common faults in systems.

Faults may also be classified according to their time behavior, a fault which appears abruptly is known as abrupt fault, a slowly growing fault is called incipient fault, a fault that appears at intermittent time intervals is called intermittent fault. The temporal classification of faults is shown in Fig. 4. The process of fault detection gives the information whether a fault has occurred or not, fault isolation pin points the source of the fault and fault identification involves reconstruction of magnitude and time behavior of the fault signal. The process of fault diagnosis includes both fault detection and fault isolation. An efficient fault detection technique should have following features.

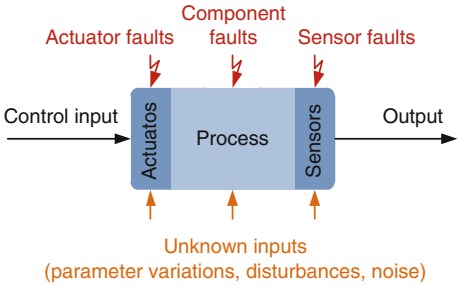


Fig. 1 Plant, which consists of plant components, sensors and actuators

- Faults should be detected as early as possible.
- Fault detection rate should be as large as possible and, simultaneously, false alarm rate should be as small as possible. Or, equivalently, the fault detection scheme should be sensitive to faults and robust against other inputs.
- Fault detection scheme should be applicable for supervision of dynamic process.
- Fault detection technique should involve less computational cost so that online implementation is possible.
- Design procedure should be simple.

The next section will give a brief description of fault diagnosis schemes.

1.3 Fault Detection Techniques

The history of fault detection is quite old; the earliest approach for fault detection was by using biological senses such as by looking for changes in color, and shape, listening to sounds unusual in pitch and loudness, touching to feel heat and vibration, and smelling for fumes because of overheating or leakage [17]. With industrial revolution, there was a need to autonomously detect faults. This not only saves extra labor but is also more reliable, quick and efficient. Furthermore, some locations may not be accessible or dangerous to human being. The simplest autonomous fault detection technique is by limit checking. By this approach, the occurrence of faults is indicated if the measured variable crosses a preset upper or lower bound. Although simple, this technique can only detect faults if they grow enough to cross the bounds, therefore, the process of fault detection by this approach is slower. Additionally, this approach cannot be efficiently applied to dynamic systems. Therefore, advanced methods of fault detection must be employed.

Figure 5 shows a classification of advanced fault detection techniques. These can be broadly grouped into signal based techniques and model-based techniques. In signal-based approaches, the information of fault is obtained by collecting some properties of measured signals. Examples of these properties are the magnitudes of the time function, trend checking from the derivative, mean and variance, spectral power densities, correlation coefficients, etc., of the measured signals. Figure 6 shows the conceptual diagram for signal-based fault detection schemes.

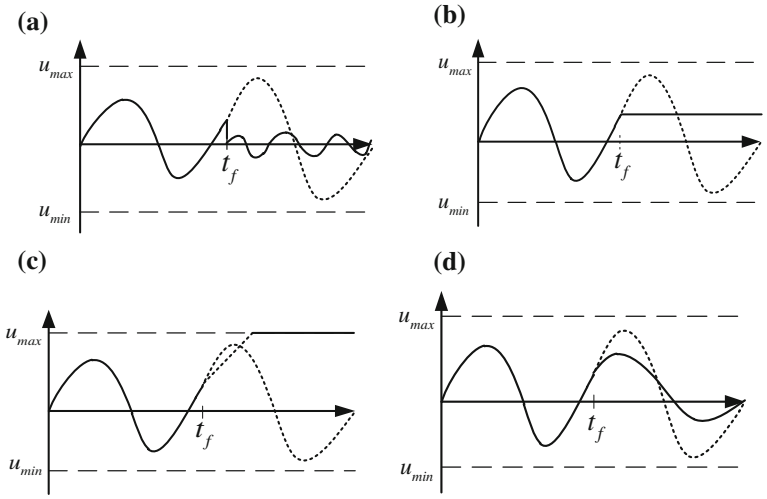


Fig. 2 Graphical representation of common types of actuator faults in servomotors. *Dotted lines* show the desired value of actuator and the *solid lines* show actual value [11]. **a** Floating around trim, **b** Lock-in-place, **c** Hard-over failure, **d** Loss of effectiveness

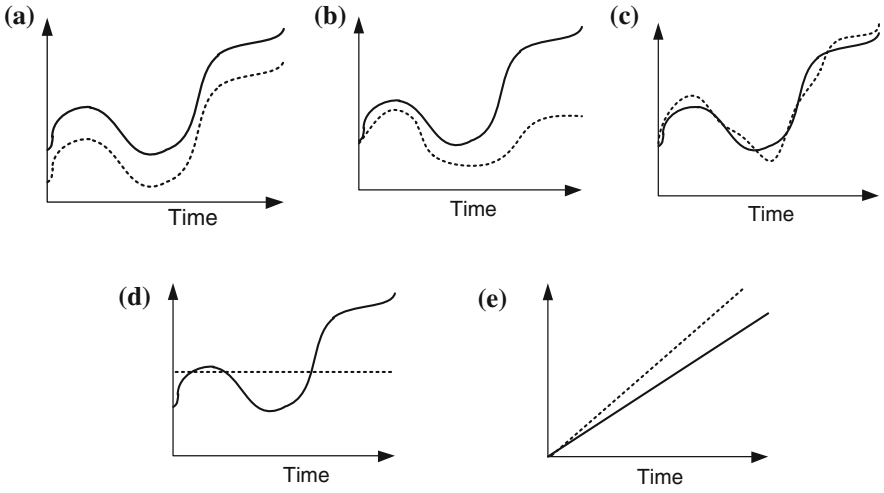


Fig. 3 Graphical depiction of different kinds of sensor faults. *Solid lines* show the actual values whereas the *dotted lines* show the measured values [5]. **a** Bias, **b** Drift, **c** Loss of accuracy, **d** Freezing, **e** Calibration error

Model-based fault detection techniques use a model of process, in addition to the measured signals, for the purpose of fault detection. The model can be an analytical model represented by set of differential equations or it can be knowledge-based

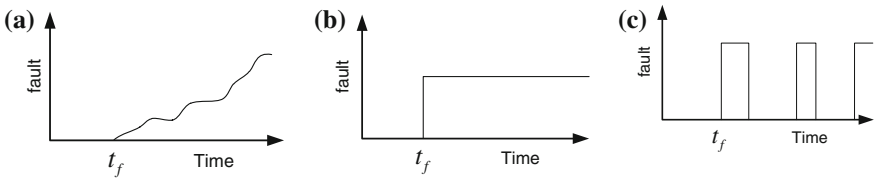


Fig. 4 Graphical illustration of abrupt, incipient and intermittent faults occurring at time t_f. **a** Abrupt fault, **b** Incipient fault, **c** Intermittent fault

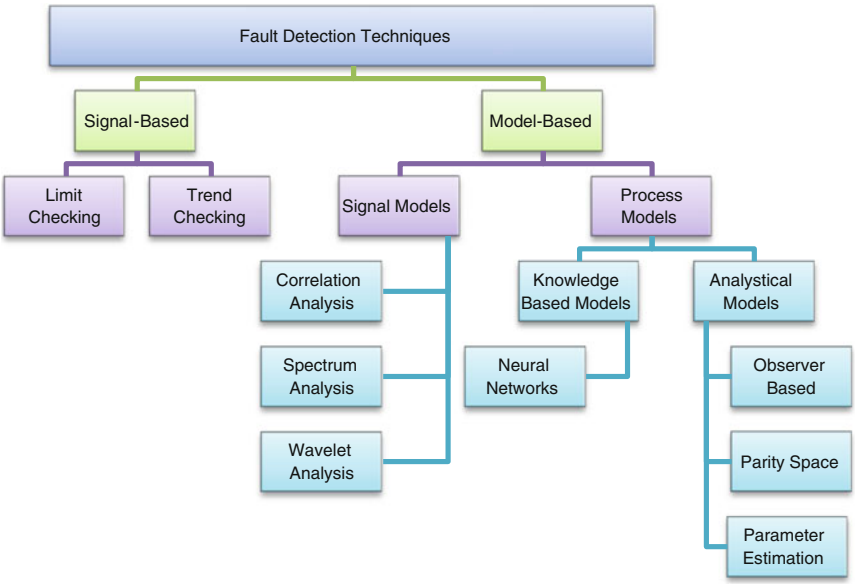


Fig. 5 Classification of fault detection schemes

model represented by, for example, neural networks, petri nets, experts systems, fuzzy rules etc. Knowledge-based model approaches do not need full analytical modeling, therefore, are more suitable in information-poor systems or in situations where the mathematical model of the process is difficult to obtain or is too complex. This is the case, for example, in chemical processes which are difficult to model analytically. Figure 7 shows the schematic diagram of a model-based fault detection scheme. It consists of two main stages; residual generation and residual evaluation. The objective of residual generation is to produce a signal, called residual signal, by comparing measurements with their estimates. The purpose of residual evaluation is to inspect the residual signal for possible presence of faults.

Analytical model based approaches include observer-based techniques, parity space approaches and parameter identification approach. The next section describes observer-based approaches for fault detection in linear systems.

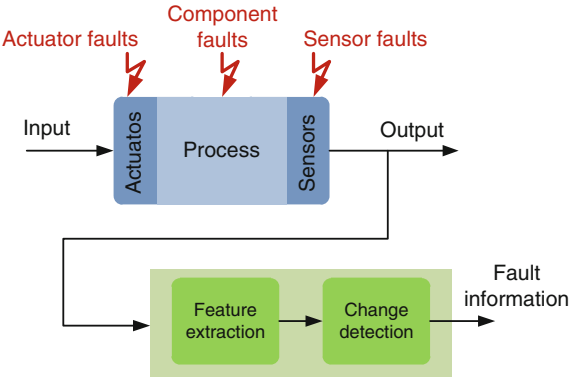


Fig. 6 Signal based fault detection

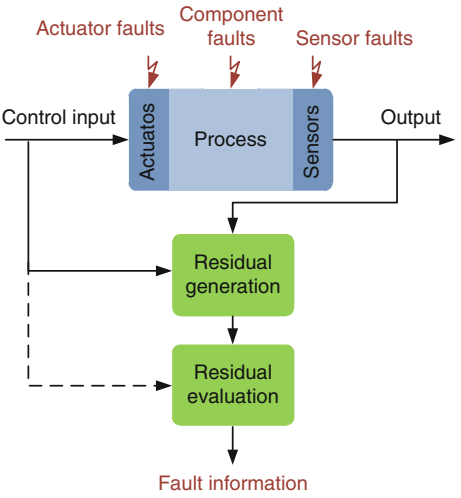


Fig. 7 Model-based fault detection

2 Observer-Based Fault Detection in Linear Systems

Fault detection in linear systems has been extensively studied and there exist well established fault detection techniques for linear systems (see for example [4, 6, 8, 15, 20, 21]). Observer-based fault detection is one of the earliest fault detection techniques; the preliminary research was presented in the celebrated papers of Beard and Jones [2, 23]. In observer-based fault detection technique, residual signal is generated by comparing the measurements with their estimates generated by an observer (filter), as shown in Fig. 8. In ideal case, that is, if mathematical model is accurate and there are no disturbances, the estimate generated by the observer matches the measurements in fault-free case and the residual signal is zero. A nonzero residual gives the indication of occurrence of fault. However, 100 % accurate mathematical modeling is either not possible or is too complex to

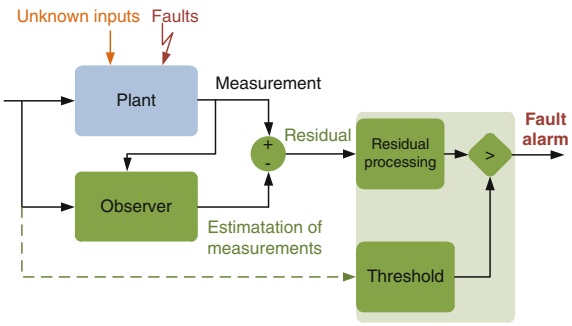


Fig. 8 Observer-based fault detection

handle. Additionally, external disturbances and measurement noise (collectively known as unknown inputs) are always present. Therefore, a threshold is computed which cares for the effect of uncertainties and unknown inputs, and if the residual crosses the threshold, it gives an indication of fault. Thus, following are two design objectives in the observer based fault detection scheme.

1. Design of an observer which gives correct estimation of measurements, that is, the generated residual signal is robust against unknown inputs and is sensitive to faults.
2. A too high threshold will result into some faults undetected and a too low threshold will result into high false alarm rate. Therefore, computation of an optimal threshold is a design objective.

There are several different structures of observers and many advanced techniques exist for the design of observers to achieve the above mentioned objectives (see for example [6, 8]). In the sequel, a very simple and commonly used observer (filter) for fault detection will be presented.

2.1 Fault Detection Filter

Consider following linear time invariant system

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) + E_f f(t) \\ y(t) &= Cx(t) + Du(t) + F_f f(t), \end{aligned} \tag{1}$$

where, $x(t) \in \Re^n$ is the state vector, $u(t) \in \Re^{k_u}$ is the input vector, $y(t) \in \Re^m$ is the measurement vector and $f(t) \in \Re^{k_f}$ is the fault vector to be detected. A, B, E_f, C, D, F_f are known matrices of appropriate dimensions. For the purpose of residual generation, we use fault detection filter which has following structure,

$$\begin{aligned}\dot{\hat{x}}(t) &= A\hat{x}(t) + Bu(t) + L(y(t) - \hat{y}(t))\\ \hat{y}(t) &= C\hat{x}(t) + Du(t)\\ r(t) &= y(t) - \hat{y}(t),\end{aligned} \tag{2}$$

where, $\hat{x}(t) \in \Re^n$ is the estimation of the state vector, $\hat{y}(t) \in \Re^m$ is the estimation of the measurement vector, and $r(t) \in \Re^m$ is the residual signal. The matrix $L \in \Re^{n\times m}$ is called observer (filter) gain matrix and is a design parameter. It is so selected/computed that $\lim_{t\to\infty} x(t) - \hat{x}(t) = 0$ for $f = 0$. Define $e(t) = x(t) - \hat{x}(t)$, then dynamics of the residual generator, in fault-free case, can be written as:

$$\begin{aligned}\dot{e}(t) &= (A - LC)e(t)\\ r(t) &= Ce(t).\end{aligned}$$

It can be seen that if the filter gain matrix L is chosen such that *A–LC* has all the eigenvalues with negative real parts, then

$$\lim_{t\to\infty} x(t) - \hat{x}(t) = 0.$$

2.2 Sensitivity to Faults and Robustness Against Disturbance

The fault detection filter presented in Sect. 2.1 is a very basic type of residual generator. It ensures that in fault-free case residual will converge to zero. The basic assumption is that there is no modeling uncertainty and no unknown inputs. However, this is ideal situation and, in general, these assumptions are never satisfied. Modeling uncertainties and unknown inputs are always there which make the residual to be non-zero even if there is no fault. To that end, consider a linear time invariant system with unknown inputs described by,

$$\begin{aligned}\dot{x}(t) &= Ax(t) + Bu(t) + E_d d(t) + E_f f(t)\\ y(t) &= Cx(t) + Du(t) + F_d d(t) + F_f f(t),\end{aligned} \tag{3}$$

where $d(t) \in \mathbb{R}^{k_d}$ is vector of unknown inputs which include both the external process disturbance and the sensor noise. Among many other possibilities, following are two approaches to handle the systems with unknown inputs.

2.2.1 Perfect Unknown Input Decoupling

The prime objective of advanced fault detection scheme is to generate residual signal which is sensitive to faults as much as possible and highly robust against process disturbances and measurement noises. A pretty natural choice is to make

the residual signal completely insensitive to unknown inputs (measurement noises and process disturbances). This problem was discussed heavily in 1970s and till mid 1980s. Interested reader is referred to [6, 8]. In what follows, we describe the problem of decoupling unknown inputs which is referred to as Perfect Unknown Input Decoupling (PUID).

Consider the system (3) and a residual generator,

$$\dot{\hat{x}}(t) = A\hat{x}(t) + Bu(t) + L(y(t) - \hat{y}(t))$$
$$r(t) = V(y(t) - \hat{y}(t)).$$

The aim is to design a residual generator by finding L, V such that the residual generator is stable and

- $vC(sI - A + LC)^{-1}Ed = 0$
- $v(C(sI - A + LC)^{-1}(E_f - LF_f) + F_f) \neq 0.$

This will ensure complete elimination of the effect of unknown inputs on the residual signal.

2.2.2 Robust Residual Generation Using Standard Robust Observer Theory

The problems associated with perfect decoupling of unknown inputs from the residual is that if faults share the same subspace as that of unknown inputs, then upon decoupling unknown inputs, fault will also be decoupled. The detection mechanism will not be able to detect it. It is worth noting that fault is a physical phenomenon and if care is not taken at earliest, may lead to disastrous implications. It is, therefore, quite natural to design a residual generator in the framework of improving robustness against unknown inputs and sensitivity to faults. To this end, various optimization indices have been proposed in literature, see for instance [6, 9, 18, 25, 46] and references therein. These optimization indices are mainly based on H_∞/H_∞, H_2/H_2, H_-/H_∞. Among them, H_-/H_∞ trade-off based FDF design, initiated in [9, 18], has received considerable attention in recent years, for instance see [25, 46] and the references therein. $H_\infty-$ norm measures robustness against the unknown inputs. In other words, it measures the effect of those unknown inputs which have maximum influence on the residual signal. Likewise, the sensitivity to faults is measured by the so-called H_- index. It is the measure of the minimum influence of the fault on the residual signal. In case of fault detection for LTI systems, H_- index represents the smallest singular value of the fault to residual transfer matrix. It can be regarded as the worst case (minimum "nonzero") fault sensitivity. Similarly, $H_\infty-$ norm represents the largest singular value of the residual to unknown input transfer matrix. Multi-objective optimization problem to simultaneously maximize the effect of fault on the residual (that is, to increase the H_- index of the residual generator), and to minimize the effect

of unknown inputs (that is, to decrease the $H_\infty-$ norm of the residual generator) has been studied and LMI formulation in presented in many research papers, interested reader is referred to [8, 25, 46].

3 Fault Detection in Nonlinear Systems

Fault detection in nonlinear system is very challenging task due to nonlinearities present in nonlinear systems. Interests in fault detection and isolation for nonlinear systems have grown significantly in recent years due to the fact that most of the systems, we face in practice, are nonlinear in nature. There exist a number of techniques for fault detection, among them; the so-called observer-based fault detection is widely studied. In addition, this technique has been proven efficient in detecting faults. Various problems related to observer-based fault detection in nonlinear systems both from practical as well as theoretical interest have been addressed. A comprehensive study of fault detection in non-linear systems can be found in [1, 16, 24]. Here, we present only simple fault detection techniques for nonlinear systems.

3.1 Linearization Techniques

Fault detection of LTI system is a mature area. It is quite natural to linearize the nonlinear system and use the linear techniques to solve the fault detection problem. Taylor series approximation is used to linearize the system around the point of interest. Consider a nonlinear system

$$\begin{aligned} \dot{x}(t) &= f(x(t), u(t)), \qquad x(0) = x_0 \\ y(t) &= h(x(t), u(t)) \end{aligned} \tag{4}$$

The nonlinear system (4) can be linearized at an operating points (x^*, u^*) using Taylor series expansion. For that purpose, define deviation variables,

$$\begin{aligned} \delta_x(t) &\equiv x(t) - x^* \\ \delta_u(t) &\equiv u(t) - u^*. \end{aligned}$$

Substitute these definitions in (4) to get

$$\begin{aligned} \dot{\delta}_x(t) &= f(x^* + \delta_x(t), u^* + \delta_u(t)) \\ y &= h(x^* + \delta_x(t), u^* + \delta_u(t)). \end{aligned}$$

By Taylor series expansion of the right hand side,

$$\dot{\delta}_x(t) = f(x^{\star}, u^{\star}) + \frac{\partial f}{\partial x}|_{\substack{x=x^*\\u=u^*}} \delta_x(t) + \frac{\partial f}{\partial u}|_{\substack{x=x^*\\u=u^*}} \delta_u(t) + HOT$$
$$y = h(x^{\star}, u^{\star}) + \frac{\partial h}{\partial x}|_{\substack{x=x^*\\u=u^*}} \delta_x(t) + \frac{\partial h}{\partial u}|_{\substack{x=x^*\\u=u^*}} \delta_u(t) + HOT,$$

where *HOT* represents second and higher order terms in the Taylor series expansion. Neglecting *HOT*, using the fact that at an equilibrium point $f(x^{\star}, u^{\star}) = 0$ and setting $A = \frac{\partial f}{\partial x}\Big|_{\substack{x=x^*\\u=u^*}}$, $B = \frac{\partial f}{\partial u}\Big|_{\substack{x=x^*\\u=u^*}}$, $C = \frac{\partial h}{\partial x}\Big|_{\substack{x=x^*\\u=u^*}}$, and $D = \frac{\partial h}{\partial u}\Big|_{\substack{x=x^*\\u=u^*}}$, the linear approximation of the nonlinear system (4) at the operating point becomes,

$$\dot{\delta}_x(t) = A\delta_x(t) + B\delta_u(t)$$
$$y = C\delta_x(t) + D\delta_u(t).$$

Note that this is an approximate model for the nonlinear process (4) and will be valid locally around the linearization point. Once the nonlinear system is linearized, developed techniques for fault detection in linear systems, for example those discussed in Sect. 2, can be applied for the generation of residual signal.

3.2 Extended Observers

The linearization technique discussed in the last section has the advantage that it is simple. However, the linearized model will not be valid if the operating region is too wide and the system is operating away from the linearization point. Therefore, a possible solution is to linearize the system at each time instant at current operating point, instead of a fixed linearization point. That is, at each time instant, there will be a linearized model and residual generator is designed at each time instant. The advantage is that the linearization error will be small. However, an associated disadvantage is additional computational cost because at each time instant residual generator has to be redesigned.

In addition to the above mentioned techniques for residual generation in nonlinear systems, there are many approaches which focus to design a residual generator directly for nonlinear systems. The direct design approaches do not fall in the scope of this book, interested reader is referred to [1, 24]. At this point it is important to emphasize that direct design approaches are either only applicable to very limited class of nonlinear systems or the design procedure involves solution of nonlinear partial differential equations, for which even the numerical solution is quite hard to obtain. This is the major motivation to study TS fuzzy approach for fault detection which is presented in the next sections.

4 TS Fuzzy Modeling of Nonlinear Systems

Consider a nonlinear system described by,

$$\begin{aligned} \dot{x}(t) &= f(x(t), u(t)) \\ y(t) &= h(x(t), u(t)). \end{aligned} \tag{5}$$

The nonlinear system is linearized at N linearization points using Taylor series approximation and each linear model has the form given by,

$$\begin{aligned} \dot{x}(t) &= A_i x(t) + B_i u(t) \\ y(t) &= C_i x(t) + D_i u(t), \qquad i = 1, 2, \cdots, N. \end{aligned} \tag{6}$$

Then the global behavior of the nonlinear system can be described by fuzzy combination of local linearized models [41]. The process of fuzzy approximation of a nonlinear system by the linearized models consists of three major steps, these are:

- Fuzzification
- Application of rule base
- Defuzzification.

To understand these steps, a few basic concepts are required which are described below.

4.1 Basics

4.1.1 Fuzzy Sets

A fuzzy set is a set without a clearly defined boundary. It is different from the classical sets in the sense that the members of fuzzy set may have partial memberships compared to the classical set which either wholly includes or excludes an element. For example, consider the following two sets,

$$\begin{aligned} A &= \{1, 2, 3\}, \\ B &= \{\text{set of low temperatures}\} \end{aligned}$$

Here, A is a classical set and it is clear that 1, 2 and 3 are members of the set whereas 6, 7 are not. B is a fuzzy set, for fuzzy set B, some elements may be partial members, for example -10 ^{o}C is definitely a member of the set but 0 ^{o}C is only partially member of the set. To define the degree of membership, a membership function is used. These membership functions may have different shapes, for example, a rectangular membership function, a triangular member-ship function, a sigmoid membership function etc. Figure 9 shows a sigmoid membership function

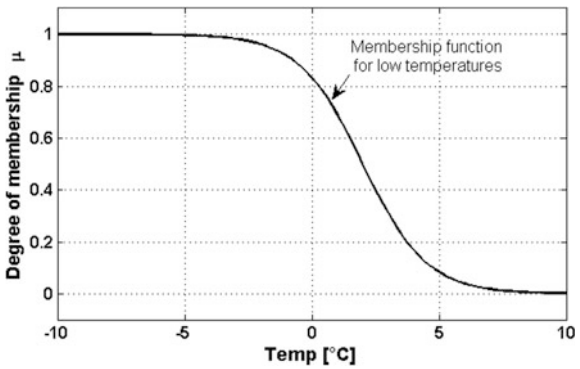


Fig. 9 Sigmoid membership function

for the set of low temperatures. The degree of membership of each value of temperature can be inferred from the membership function, it can be seen that − 10 ^{o}C has 100 % membership with the set, whereas 0 ^{o}C has 80 % membership with the set.

4.1.2 Logical Operations

The logical operations used in fuzzy logic are similar to the logical operations in Boolean algebra. Table 1 shows the Boolean logical operations. Similar to 'AND' and 'OR' operations in Boolean algebra, we can define 'AND' and 'OR' operations for fuzzy variables. There are different approaches to define these operations, for example, the 'AND' operation for fuzzy variables may be defined as the minimum of the two operands, or it is sometimes also defined as the product of the two operand. In this study, we shall take the product of the operand as fuzzy AND operation. With this basic knowledge, now we describe the steps of TS fuzzy modeling of nonlinear systems.

4.2 Fuzzification

Fuzzification is the process of transforming crisp values into linguistic variables, that is, determining the grade of membership of some quantity to a fuzzy set. For that purpose, membership functions are used. The procedure for Fuzzification is illustrated in Fig. 10; a variable $z(t)$ which has a crisp value z_0 is fuzzified using triangular membership functions. It can be seen that z_0 has 80 % membership with the set M_1 and 40 % membership with set M_2. The notation $M_1(z(t))$ means the Fuzzification of $z(t)$ using the membership function M_1.

Table 1 Boolean logical operations

A	B	A and B	A or B
0	0	0	0
0	1	0	1
1	0	0	1
1	1	1	1

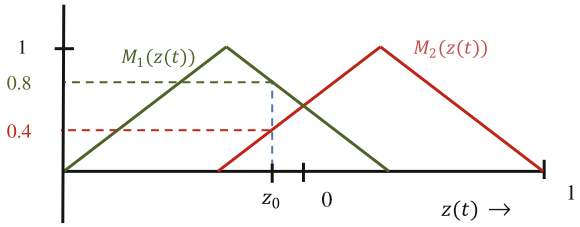


Fig. 10 Fuzzification of a crisp value z_0 of a variable $z(t)$

4.3 IF–THEN Rules

IF–THEN rules specify the core of fuzzy logic. For the nonlinear system (5) and the corresponding linearized models (6), the IF–THEN rules are stated as,

Rule i ($i = 1, 2, \cdots, N$)

IF $z_1(t)$ is M_{i1} AND $\cdots$ AND $z_\ell(t)$ is $M_{i\ell}$ THEN

$$\dot{x}(t) = A_i x(t) + B_i u(t)$$
$$y(t) = C_i x(t) + D_i u(t).$$

The rule has two distinct parts; the IF part of the rule is called the antecedent or premise and the THEN part is called the consequent. In the above rule, $z_i(t)$ is called the premise variables and is a known parameter. It may depend upon states or inputs of the system; generally, the nonlinear part in the mathematical model of the system is taken as the premise variable. M_{ij} represents the membership function of the fuzzy sets and N is the number of rules. To evaluate the antecedent, the premise variables are first fuzzified, then fuzzy AND operation is performed, that is, the antecedent is,

$$h_i(z(t)) = \prod_{j=1}^{\ell} M_{ij}(z_j(t)). \tag{7}$$

4.4 Defuzzification

Each rule has its own implication; defuzzification is the process of aggregation of the implication of all the rules. The most commonly used aggregation method is

center of gravity. Under the center of gravity method of defuzzification, the global behavior of the nonlinear system in terms of linearized sub-models is given as,

$$\begin{aligned} \dot{x}(t) &= \sum_{i=1}^{N} \mu_i(z(t))[A_i x(t) + B_i u(t)] \\ y(t) &= \sum_{i=1}^{N} \mu_i(z(t))[C_i x(t) + D_i u(t)], \end{aligned} \tag{8}$$

where

$$\mu_i(z(t)) = \frac{h_i(z(t))}{\sum_{i=1}^{N} h_i(z(t))}. \tag{9}$$

4.5 Example

To demonstrate the fuzzy modeling of nonlinear systems, we take following nonlinear system [43]

$$\dot{x}(t) = \begin{bmatrix} -1 & x_1(t)x_2^2(t) \\ 3 + x_1^2(t)x_2(t) & -1 \end{bmatrix} x(t), \tag{10}$$

where $x(t) = [\, x_1(t) \quad x_2(t) \,]^T$ and $x_1(t)x_2^2(t)$, $3 + x_1^2(t)x_2(t)$ are nonlinear terms. These nonlinear terms are taken as premise variables, that is $z_1(t) = x_1(t)x_2^2(t)$ and $z_2(t) = 3 + x_1^2(t)x_2(t)$. Then, we have

$$\dot{x}(t) = \begin{bmatrix} -1 & z_1(t) \\ z_2(t) & -1 \end{bmatrix} x(t).$$

For simplicity, it is assumed that $x_1(t) \in [-1\ 1]$ and $x_2(t) \in [-1\ 1]$ which in turn results into $z_1(t) \in [-1\ 1]$ and $z_2(t) \in [0\ 4]$. For the premise variables, we take the following membership functions

$$M_{11}(z_1(t)) = \frac{1 + z_1(t)}{2}, \quad M_{21}(z_1(t)) = \frac{1 - z_1(t)}{2},$$
$$M_{12}(z_2(t)) = \frac{z_2(t)}{4}, \qquad M_{22}(z_2(t)) = \frac{4 - z_2(t)}{4}.$$

These membership functions are shown in Fig. 11. The nonlinear system is linearized at four linearization points; the linearized sub-models are given as,

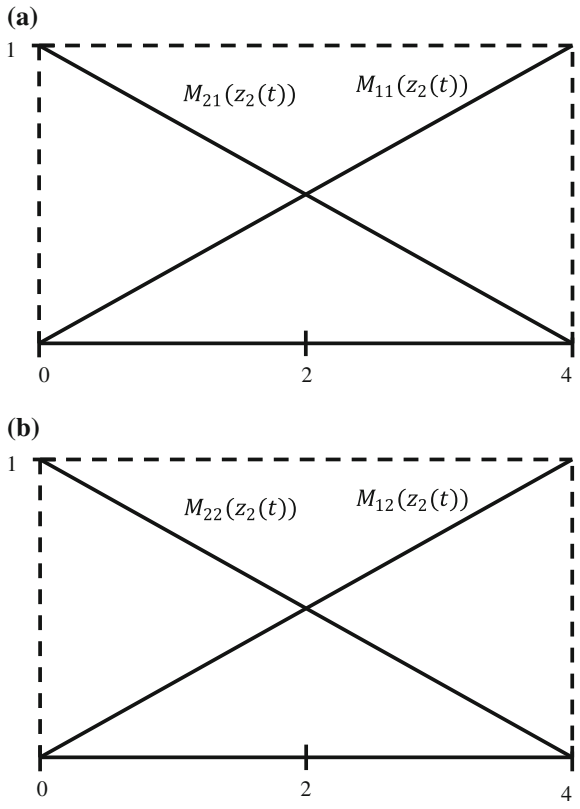


Fig. 11 Membership functions for the premise variables $z_1(t)$ and $z_2(t)$

$$\dot{x}(t) \quad = A_i x(t), \qquad i = 1, 2, 3, 4 \tag{11}$$

and the matrices A_i come out to be

$$A_1 = \begin{bmatrix} -1 & 1 \\ 4 & -1 \end{bmatrix}, \quad A_2 = \begin{bmatrix} -1 & 1 \\ 0 & -1 \end{bmatrix},$$
$$A_3 = \begin{bmatrix} -1 & -1 \\ 4 & -1 \end{bmatrix}, \quad A_4 = \begin{bmatrix} -1 & -1 \\ 0 & -1 \end{bmatrix}.$$

Now, according to (8), the global behavior of the nonlinear system is described by fuzzy combination of the linearized sub-models, that is,

$$\dot{x}(t) = \sum_{i=1}^{4} \mu_i(z(t)) A_i x(t), \tag{12}$$

where, according to (7),

$$\begin{aligned} h_1(z(t)) &= M_{11}(z_1(t))M_{12}(z_2(t)), \\ h_2(z(t)) &= M_{11}(z_1(t))M_{22}(z_2(t)), \\ h_3(z(t)) &= M_{21}(z_1(t))M_{12}(z_2(t)), \\ h_4(z(t)) &= M_{22}(z_1(t))M_{22}(z_2(t)), \end{aligned}$$

and $\mu_i(z(t))$ is as described in (9). It should be emphasized that the fuzzy model described by (12) exactly describes the nonlinear system (10) in the interval $x_1(t) \in [-1\ 1]$ and $x_2(t) \in [-1\ 1]$.

5 TS Fuzzy Approach for Fault Detection

Different techniques for fault detection in nonlinear systems were presented in Sect. 3. Among these techniques, the linearization approach presented in Sect. 3.1 is simple and easy to apply. However it is only applicable if the operating region is small and nonlinearity is not too high. For the systems with wide operating region or with high nonlinearities, the nonlinear fault diagnosis techniques should be applied. But these techniques are either limited to a very small set of nonlinear systems or the design procedure involves solution of nonlinear partial differential equations, for which even the numerical solution is quite difficult to obtain. TS Fuzzy approach is an alternate solution. With this approach, the nonlinear system is linearized at multiple points, observers are designed for local linear models using linear techniques and the global state/output estimation is obtained by fuzzy fusion of local observer output, as illustrated in Fig. 12.

Consider again the nonlinear system given by (5) with external disturbances and faults,

$$\begin{aligned} \dot{x} &= f(x,u) + g_f(x)f + g_d(x)d \\ y &= h(x,u). \end{aligned}$$

The equivalent TS fuzzy models for the nonlinear system is described by

$$\begin{aligned} \dot{x} &= \sum_{i=1}^{N} \mu_i(z(t))[A_i x + B_i u + E_{d,i} d + E_{f,i} f] \\ y &= \sum_{i=1}^{N} \mu_i(z(t))[C_i x + D_i u + F_{d,i} d + F_{f,i} f] \end{aligned}$$

where $\mu_i(z(t))$ as defined in (9). For the purpose of residual generation, a fault detection filter is designed for each sub-model and the overall state estimation is the fuzzy combination of individual local observers, that is

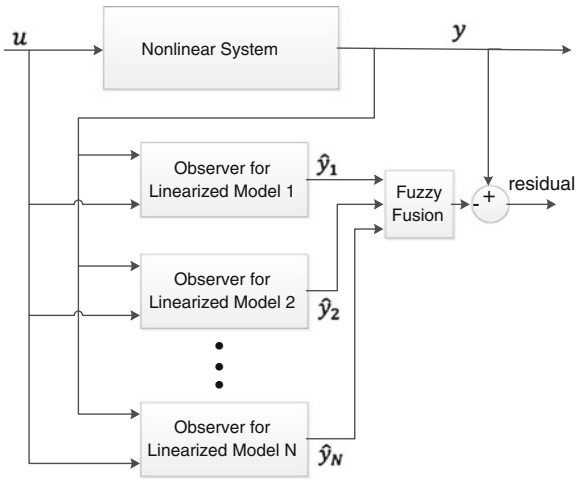


Fig. 12 Fuzzy observer for residual generation

$$\dot{\hat{x}} = \sum_{i=1}^{N} \mu_i(z(t))[A_i\hat{x} + B_i u + L(y - \hat{y})]$$
$$\hat{y} = \sum_{i=1}^{N} \mu_i(z(t))[C_i\hat{x} + D_i u]$$
$$r = y - \hat{y}.$$

The most important requirement of residual generator is stability. To discuss stability of TS Fuzzy residual generator, the error dynamics will be analyzed. Define $e = x - \hat{x}$, the dynamics of the residual generator is given by

$$\dot{e} = \sum_{i=1}^{N}\sum_{j=1}^{N} \mu_i(z(t))\mu_j(z(t))((A_i - L_iC_j)e + (E_{d,i} - L_iF_{d,j})d + (E_{f,i} - L_iF_{f,j})f)$$
$$r = Ce + F_{d,i}d + F_{f,i}f.$$

As stated earlier, a residual must be zero in the absence of faults and external disturbances. The dynamics of the residual generator with f = 0 and d = 0 are given by,

$$\dot{e} = \sum_{i=1}^{N}\sum_{j=1}^{N} \mu_i(z(t))\mu_j(z(t))[(A_i - L_iC_j)e] \quad (13)$$
$$r = Ce.$$

It is obvious from (13) that the residual will converge to zero if and only if the error dynamics are asymptotically stable. To this end, the following theorem is useful.

Theorem 1. [44] *The residual generator in (13) is asymptotically stable if there exists a common positive definite matrix P such that;*

$$(A_i - L_i C_i)^T P + P(A_i - L_i C_i) < 0 \qquad \forall\, i = 1, 2, \cdots, N$$

and

$$\left(\frac{A_i - L_i C_i + A_j - L_j C_j}{2}\right)^T P + P\left(\frac{A_i - L_i C_i + A_j - L_j C_j}{2}\right) < 0, \quad i < j \leq N$$

This means that a common P must exist for all the sub-models [42, 43]. If the system does not have a nonlinearity in the output equation and there is no input appearing in the output, that is,

$$y = Cx,$$

the expression for fuzzy observer is simplified, and we have

$$\begin{aligned} \dot{\hat{x}} &= \sum_{i=1}^{N} \mu_i(z(t))[A_i \hat{x} + B_i u + L(y - \hat{y})] \\ \hat{y} &= C\hat{x}. \end{aligned} \tag{14}$$

Furthermore, the residual generator dynamics are given by,

$$\begin{aligned} \dot{e} &= \sum_{i=1}^{N} \mu_i(z(t))[(A_i - L_i C_j)e] \\ r &= Ce. \end{aligned}$$

The stability conditions for the observer in this special case are also simplified, which are given by the following corollary.

Corollary 1. [38] *The simplified observer described by (14) is asymptotically stable if there exists a common positive definite matrix P such that*

$$(A_l - L_i C_i)^T P + P(A_l - L_i C_i) < 0 \qquad \forall i = 1, 2, \cdots, N.$$

Thus, with the TS Fuzzy approach, the design of residual generator for non-linear systems transforms into finding stable observers for linearized models. It is to emphasize that the theory to design observers for linear systems is well established and can be found in many standard text books (see for example [6, 8]).

5.1 Example

To demonstrate the application of TS Fuzzy model for fault detection of non-linear systems, we design a fault detection filter for the the nonlinear system presented in Example 1.4.5. We assume that the state $x_1(t)$ is measurable, that is,

$$y(t) = x_1(t).$$

The equivalent fuzzy model for the system was described by (12). For each of the linearized models in (11), a fault detection filter of the form (2) is designed, the observer gains come out to be,

$$L_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}, L_2 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}, L_3 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}, L_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix},$$

which place the eigenvalues of each of the observer at -2 and -3. Then the global state estimation is generated by

$$\begin{aligned} \dot{\hat{x}}(t) &= \sum_{i=1}^{4} \mu_i(z(t))[A_i x(t) + L_i(y - \hat{y})] \\ \hat{y}(t) &= C\hat{x} \end{aligned}$$

and the residual signal can be generated by

$$r(t) = y(t) - \hat{y}(t).$$

6 Robust Residual Generation with TS Fuzzy Models

In the last section, a simple fuzzy observer design procedure for nonlinear dynamical systems was described. As we know that, for successful fault detection, the residual signal must satisfy the properties listed in Sect. 1, these desired features are rewritten here,

- The residual signal should converge to zero in the absence of faults and disturbances, that is,

$$\lim_{\substack{t \to \infty \\ d=0, f=0}} r(t) = 0.$$

- The residual signal should as much sensitive to faults as possible so that the faults are successfully detected. This requirement stated in terms of L_2 gain is

$$\inf_{f \neq 0, d=0} \frac{\|r(t)\|_2}{\|f(t)\|_2} \geq \beta.$$

- The residual signal should be robust against disturbances, that is

$$\sup_{d \neq 0, f=0} \frac{\|r(t)\|_2}{\|d(t)\|_2} \leq \gamma.$$

Last section described a simple procedure to design residual generator for TS fuzzy models of nonlinear systems. However, that observer guaranteed only the first one among the above mentioned features of a residual signal. This was achieved by selecting an observer gain matrix L_i such that the eigenvalues of $A_i - L_iC$ have negative real part. However, the particular choice of the eigenvalues is purely arbitrary. It means that there is an additional degree of freedom in the choice of the observer gain matrices. This additional degree of freedom can be utilized to achieve the rest of the two design objectives. A study of the design procedure to achieve the disturbance robustness and fault sensitivity features of TS fuzzy observer-based residual generator can be found in [12–14].

7 Fault Detection of TS Fuzzy Systems with Immeasurable Premise Variables

Section 5 described an approach for residual generation for nonlinear systems using TS Fuzzy approach. The beauty of TS Fuzzy approach lies in the fact that the global behavior of a nonlinear system can be exactly described by fuzzy combination of local linearized models. Therefore, the residual generation techniques for linear systems can be quickly applied to nonlinear systems. However, in the above study, the basic assumption was that the premise variables are measurable/known. Premise variables are generally some known function of states of the system. In most of real system, all the states of the system are either not measurable or are intentionally not measured to avoid additional cost. Therefore, the states are (partially) unknown. In that case, the only choice is to use the estimate of the states, that is the observer is given by,

$$\begin{aligned}\dot{\hat{x}} &= \sum_{i=1}^{N} \mu_i(\hat{z}(t))[A_i\hat{x} + B_iu + L(y - \hat{y})] \\ \hat{y} &= \sum_{i=1}^{N} \mu_i(\hat{z}(t))[C_i\hat{x} + D_iu] \\ r &= y - \hat{y},\end{aligned}$$

where $\hat{z}(t) \equiv z(\hat{x}(t))$. In this situation, state estimation error will no more be governed by (13). This means that the analysis carried out in Sect. 5 will be no more valid. The problem of residual generation for systems with immeasurable premise variables has been discussed in [19]. To simplify the discussion, let us take a system with no nonlinearity and no input in output equation,

$$\begin{aligned}\dot{x}(t) &= \sum_{i=1}^{N} \mu_i(x(t))[A_ix(t) + B_iu(t)] \\ y(t) &= Cx(t).\end{aligned} \tag{15}$$

The observer dynamics with immeasurable premise variables are represented by,

$$\begin{aligned}\dot{\hat{x}}(t) &= \sum_{i=1}^{N} \mu_i(\hat{x}(t))[A_i x(t) + B_i u(t) + L_i(y(t) - \hat{y}(t))] \\ \hat{y}(t) &= C\hat{x}(t).\end{aligned} \tag{16}$$

Then the observer error dynamics are given by,

$$\dot{e}(t) = \sum_{i=1}^{N} \mu_i(\hat{x}(t))[A_i - L_i C]e(t) + \Delta(x(t), \hat{x}(t), u(t))$$

with

$$\Delta(x(t), \hat{x}(t), u(t)) = \sum_{i=1}^{N} (\mu_i(x(t)) - \mu_i(\hat{x}(t)))[A_i x(t) + B_i u(t)]$$

and

$$\|\Delta(x(t), \hat{x}(t), u(t))\| \leq \|x - \hat{x}\|$$

For fuzzy systems with immeasurable premise variables, the following Theorem is useful.

Theorem 2. [3] *The state estimation error between the multiple model (15) and the multiple observer (16) converges globally asymptotically toward zero, if there exist matrices* $P > 0$ *and* $Q > 0$ *such that the following conditions hold for* $i = 1,\cdots, N$

$$A_i^T P + PA_i - C^T Y_i^T - Y_i C < -Q$$

$$\begin{bmatrix} -Q + \alpha^2 I & P \\ P & -I \end{bmatrix} < 0$$

The observer gains are given by

$$L_i = P^{-1} Y_i$$

The above theorem presents an approach to design the local observers' gains. However, if the input or the states of the system are large, the term Δ gets large which in turn results into large value of α. This will produce conservative results. Further studies are carried out in [19] to produce less conservative approach for the design of the observers.

8 Application to Cyber Physical Systems

Cyber physical systems (CPS) are physical systems with integrated computational and communicational resources. Many modern systems are emerging as cyber physical systems and CPS find its application in areas as diverse as automotive industry, chemical processes, transportation, energy, aerospace, etc. Some very common examples of CPS are power generation and distribution networks, unmanned aerial vehicles and SCADA systems.

As stated earlier, physical systems are prone to faults, cyber physical systems are no exception. These systems are also composed of process components, actuators and sensors; therefore, the appearance of faults is unavoidable. In addition, cyber physical systems are also susceptible to cyber-attacks [31]. Therefore, there is need to detect faults/attacks on cyber physical systems. Fault detection in cyber physical systems has been discussed in a few research articles [34–36, 40, 45].

Most of the cyber physical systems can be modeled by set of differential and algebraic equations [34, 35], such systems are called descriptor systems. A nonlinear descriptor system is given by

$$\begin{aligned} E\dot{x}(t) &= f(x(t), u(t)), \qquad x(0) = x_0 \\ y(t) &= h(x(t), u(t)) \end{aligned}$$

where $E \in \mathbb{R}^{n \times n}$ is a singular matrix. The nonlinear descriptor systems with faults/attacks and disturbances is described by

$$\begin{aligned} E\dot{x} &= f(x, u) + g_f(x)f + g_d(x)d \\ y &= h(x, u). \end{aligned}$$

The procedure to detect faults in cyber physical systems is the same as for classical control system. That is, with observer-based approach, an observer is used to generate estimation of the measurements and the difference between the measurements and their estimates produces residual signal which gives the information of fault. The design of observer for nonlinear descriptor systems has been studies in literature (see [10, 28, 39, 48] and references therein). The nonlinear observer design approach for descriptor systems suffer from the same problems as discussed in Sect. 3. To avoid these limitations, TS fuzzy approach may be utilized.

TS fuzzy model for descriptor system is given by

$$\begin{aligned} E\dot{x} &= \sum_{i=1}^{N} \mu_i(z(t))[A_i x + B_i u + E_{d,i} d + E_{f,i} f] \\ y &= \sum_{i=1}^{N} \mu_i(z(t))[C_i x + D_i u + F_{d,i} d + F_{f,i} f]. \end{aligned}$$

The corresponding TS fuzzy state observer and residual generator has the form

$$E\dot{\hat{x}} = \sum_{i=1}^{N} \mu_i(w(t))[A_i\hat{x} + B_i u + L(y - \hat{y})]$$
$$\hat{y} = \sum_{i=1}^{N} \mu_i(w(t))[C_i\hat{x} + D_i u]$$
$$r = y - \hat{y}.$$

Thus, local observer for linear descriptor systems may be designed and the overall behavior is obtained by fuzzy fusion of local behavior. Interested readers may refer to [7, 34, 35, 47] for detailed study on observer design for linear descriptor systems.

9 Future Directions

In [12], LMI approach for the design of residual generator using TS Fuzzy models is proposed. The proposed approach guarantees simultaneous robustness against disturbances and sensitivity to faults. The results of [12] are extended to uncertain nonlinear systems in [14]. However, in these approaches, it is assumed that the premise variables are measurable. As we known that the premise variables are function of states, this means that we need additional sensors to measure all the states, which increases the cost of the systems and sometimes is impossible to measure some state variables (some state variables may not correspond to some physical quantity, or their location may not be accessible). A brief overview of dealing with the problem of immeasurable premise variables is presented in last section. A possible future direction of research is to propose LMI approach for simultaneous robustness against disturbances and sensitivity to the faults with immeasurable premise variables in TS Fuzzy systems. The extension to cyber physical systems (descriptor systems) with immeasurable premise variables is another direction for future work.

References

1. M, Abid M, Fault detection in nonlinear systems: an observer-based approach. Ph.D. dissertation/University of Duisburg-Essen.- Duisburg, Germany: [s.n.], 2010
2. R.V. Beard, Failure accommodation in linear systems through self-reorganization. Ph.D. dissertation/Massachusetts Institute of Technology, 1971
3. P. Bergsten, R. Palm, Thau-Luenberger observers for TS fuzzy systems. The Ninth IEEE International Conference on Fuzzy Systems, vol. 2, pp. 671–676, 2000
4. M. Blanke et al., *Diagnosis and Fault-Tolerant Control*. [s.l.]: (Springer, Berlin, Heidelberg, 2006)

5. J.D. Boskovic, R.K. Mehra, Stable adaptive multiple model-based control design for accommodation of sensor failures. American Control Conference (IEEE, Anchorage, AK, 2002), pp. 2046–2051
6. J. Chen, R.J. Patton, *Robust Model-Based Fault Diagnosis for Dynamic Systems*. [s.l.]: (Kluwer Academic Publishers, Boston, 1999)
7. M. Darouach, M. Boutayeb, Design of observers for descriptor systems. IEEE Trans. Autom. Control [s.l.]: **40**(7), 1323–1327, 1995
8. S.X. Ding, *Model-Based Fault Diagnosis Techniques—Design Schemes, Algorithms And Tools*. [s.l.]: (Springer, Berlin, 2008)
9. X. Ding, L. Guo, P.M. Frank, A frequency domain approach to fault detection of uncertain dynamic systems. Control and Decision Conference. San Antonio, Taxas: [s.n.], pp. 1722–1727, 1993
10. G.-R. Duan, D. Howe, R.J. Patton, Robust fault detection in descriptor linear systems via generalized unknown input observers. Int. J. Syst. Sci. ([s.l.]: Taylor and Francis) **33**(5), 369–377 (2002)
11. G.J.J. Ducard, in *Fault-Tolerant Flight Control and Guidance Systems: Practical Methods for Small Unmanned Aerial*, ed. by M.J. Grimble, M.A. Johnson. [s.l.]: (Springer, Berlin, 2009)
12. M.G. El-ghatwary, S.X. Ding, Z. Gao, Robust fuzzy fault detection for continuous-time nonlinear dynamic systems. Fault Detection, Supervision and Safety of Technical Processes, p. 240, 2006
13. M. El-ghatwary, Robust fuzzy observer-based fault detection for nonlinear systems. Ph.D. dissertation/Institute of Automatic Control and Complex Systems, University of Duisburg-Essen, Germany, 2007
14. M.G. El-Ghatwary, S.X. Ding, Z. Gao, Robust fault detection for uncertain takagi-sugeno fuzzy systems with parametric uncertainty and process disturbances. Proceedings of the IFAC Symposium Safeprocess, pp. 787–792, 2006
15. P.M. Frank, Fault diagnosis in dynamic systems using analytical and knowledge-based redundancy-A survey and some new results. Automatica **26**, 459–474 (1990)
16. P.M. Frank, On-line fault detection in uncertain nonlinear systems using diagnostic observers: a survey. Int. J. Syst. Sci. **25**(12), 2129–2154 (1994)
17. J.J. Gertler, *Fault Detection and Diagnosis in Engineering Systems*. [s.l.]: (Marcel Dekker, New York, 1999)
18. M. Hou, R.J. Patton, An LMI approach to H_/H_infty fault detection observers. Proceedings of the UKACC International Conference on Control, pp. 305–310, 1996
19. D. Ichalal et al., Design of observers for Takagi-Sugeno systems with immeasurable premise variables: an L2 approach. 17th IFAC World Congress, 2008
20. R. Isermann, *Fault Diagnosis Applications: Model-Based Condition Monitoring: Actuators, Drives, Machinery, Plants, Sensors and Fault Tolerant Systems*. [s.l.]: (Springer, Berlin, 2011)
21. R. Isermann, *Fault-Diagnosis Systems: An Introduction form Fault Detection to Fault Tolerance*. [s.l.]: (Springer, Berlin, 2006)
22. B. Jiang et al., Adaptive fault-tolerant tracking control of near-space vehicle using Takagi–Sugeno fuzzy models. IEEE Trans. Fuzzy Syst. [s.l.]: **18**(5), 1000–1007 (2010)
23. H.L. Jones, Failure detection in linear systems. Ph.D. dissertation/Massachusetts Institute of Technology, 1973
24. A.Q. Khan, Observer-based Fault Detection in Nonlinear Systems. Ph.D. dissertation/ University of Duisburg-Essen. Duisburg, Germany: [s.n.], 2011
25. J. Liu, J.L. Wang, G. Yang, An LMI approach to minimum sensitivity analysis with application to fault detection. Automatica **41**, 1995–2004 (2005)
26. C.J. Lopez-Toribio, R.J. Patton, S. Daley, Takagi–Sugeno fuzzy fault-tolerant control of an induction motor. Neural Comput. Appl. ([s.l.]: Springer) **9**(1), 19–28 (2000)
27. J.M. Maciejowski, C.N. Jones, MPC fault-tolerant flight control case study: flight 1862. IFAC Safe Process. Washington, USA: IFAC, pp. 121–126, 2003

28. B. Marx, D. Koenig, J. Ragot, Design of observers for Takagi-Sugeno descriptor systems with unknown inputs and application to fault diagnosis. Control Theory Appl. ([s.l.]: IET) **1**(5), 1487–1495 (2007)
29. I.S. Mofizul, T. Wu, G. Ledwich, A novel fuzzy logic approach to transformer fault diagnosis. IEEE Trans. Dielectr. Electr. Insul. [s.l.]: **7**(2), 177–186 (2000)
30. J. Mogford et al., Fatal Accident Investigation Report, Isomerization Unit Explosion Final Report, Texas City, Texas, USA, vol. 9, p. 2005, Dec 2005
31. C. Neuman, Challenges in security for cyber-physical systems. DHS Workshop on Future Directions in Cyber-Physical Systems Security, 2009
32. S.C. Olteanu et al., Fuel cell diagnosis using Takagi-Sugeno observer approach. International Conference on Renewable Energy for Developing Countries (REDEC), pp. 1–7, 2012
33. V. Palade et al., Fault diagnosis of an industrial gas turbine using neuro-fuzzy methods. Proceedings of the 15th IFAC World Congress, pp. 21–26, 2002
34. F. Pasqualetti, F. Dorfler, F. Bullo, Attack detection and identification in cyber-physical systems–part I: models and fundamental limitations. arXiv preprint arXiv:1202.6144 (2012a)
35. F. Pasqualetti, F. Dorfler, F. Bullo, Attack detection and identification in cyber-physical systems–part II: centralized and distributed monitor design. arXiv preprint arXiv:1202.6049 (2012b)
36. F. Pasqualetti, F. Dorfler, F. Bullo, Cyber-physical attacks in power networks: models, fundamental limitations and monitor design. 50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC), pp. 2195–2201, 2011
37. R.J. Patton, Fault-tolerant control: the 1997 situation. IFAC Safe Process (IFAC, Kingston Upon Hull, UK, 1997), pp. 1029–1051
38. R.J. Patton, J. Chen, C.J. Lopez-Toribio, Fuzzy observers for nonlinear dynamic systems fault diagnosis. Proceedings of the 37th IEEE Conference on Decision and Control, vol. 1, pp. 84–89, 1998
39. D.N. Shields, Observer design and detection for nonlinear descriptor systems. Int. J. Control ([s.l.]: Taylor and Francis) **67**(2), 153–168 (1997)
40. N. Srivastava, J. Srivastava, A hybrid-logic approach towards fault detection in complex cyber-physical systems. Proceedings of the Annual Conference of the Prognostics and Health Management Society, Portland, Oregon, USA, pp. 13–16, 2010
41. T. Takagi, M. Sugeno, Fuzzy identification of systems and its applications to modeling and control. IEEE Trans. Syst. Man Cybern. [s.l.]: **1**, 116–132 (1985)
42. K. Tanaka, M. Sugeno, Stability analysis and design of fuzzy control systems. Fuzzy Sets Syst. ([s.l.]: Elsevier) **45**(2), 135–156 (1992)
43. K. Tanaka, H. Wang, *Fuzzy Control Systems Design and Analysis: A Linear Matrix Inequality Approach*. [s.l.]: Wiley-Interscience, 2004
44. K. Tanaka, T. Ikeda, H. Wang, Robust stabilization of a class of uncertain nonlinear systems via fuzzy control: quadratic stabilizability, h-infty control theory, and linear matrix inequalities. IEEE Trans. Fuzzy Syst. [s.l.]: **4**(1), 1–13 (1996)
45. M. Tantawy Ashraf, Model-based detection in cyber-physical systems. Ph.D. dissertation, Vanderbilt University, 2011
46. J.L. Wang, G.H. Yang, J. Liu, An LMI approach to h-index and mixed h-/h-infty fault detection observer design. Automatica **43**, 1656–1665 (2007)
47. C.-W. Yang, H.-L.Tan, Observer design for singular systems with unknown inputs. Int. J. Control ([s.l.]: Taylor and Francis) **49**(6), 1937–1946 (1989)
48. G. Zimmer, J. Meier, On observing nonlinear descriptor systems. Syst. Control Lett. ([s.1]: Elsevier) **32**(1), 43–48 (1997)

Novel Computational Intelligence for Optimizing Cyber Physical Pre-evaluation System

Bo Xing

Abstract Owing to the quality heterogeneity of returned used products, firms engaged in re-manufacturing activities are obliged to employ 100 % inspection of these products to evaluate their quality and suitability for re-manufacturing. In addition to visual inspection, a recent tendency is to use data recorded in electronic devices (e.g., radio frequency identification (RFID)) implanted in the products. In this way, information is obtained quickly without the need for complete (and expensive) product disassembly. Nevertheless, making sense of RFID data in a complex cyber physical system (CPS) environment (which involves such as cloud computing for used product life cycle information retrieval and physically used products scanning) is a complex task. For instance, if an RFID readers fails, there may be missing values exist. The purpose of this chapter is to employ two computational intelligence (CI) optimization methods which can improve the reliability of such inspection process.

Keywords Re-manufacturability · Cyber physical pre-evaluation system · Reliability-redundancy allocation problem · Firefly algorithm · Teaching–learning-based optimization · Radio frequency identification

1 Introduction

Returns acquired by a re-manufacturer are typically highly variable in quality [16]. A significant consequence of this uncertainty is the inclusion of a collection/classification stage and a corresponding system of quality-dependent routing of supply in a reverse logistics network [13]. The potential value of sorting and

B. Xing (✉)
Department of Mechanical and Aeronautical Engineering, Faculty of Engineering, Built Environment and Information Technology, University of Pretoria, Pretoria, South Africa
e-mail: bxing2009@gmail.com

Z. H. Khan et al. (eds.), *Computational Intelligence for Decision Support in Cyber-Physical Systems*, Studies in Computational Intelligence 540,
DOI: 10.1007/978-981-4585-36-1_15,

classification product returns has been explored by different researchers such as [7] and several sorting policies have also been proposed in the literature (e.g., [4]). In addition, the management of product return is characterized also by the lack of information associated with such used products [29]. The recent emergence of networked radio frequency identification (RFID) system is a means of connecting a product tagged with an RFID chip to a network and thereby carrying complete information associated with it throughout its lifecycle. In this way information is obtained quickly without the need for complete (and expensive) product disassembly [53]. Several authors (e.g., [14, 27, 34]) have mentioned the use or potential use of RFID and related technology in the reverse logistics network. Nevertheless, making sense of RFID data is a complex task. For instance, if an RFID readers fails, there may be missing values exist [10]. The purpose of this chapter is to employ two innovative computational intelligence (CI) approaches for improving the reliability of such classification/inspection process.

The remainder of this chapter is organized as follows. Subsequent to the introduction in Sect. 1, the background of cyber physical Re-manufacturability pre-evaluation system is briefed in Sect. 2. Then, the problem statement is presented in Sect. 3 which is followed by a problem formulation detailed in Sect. 4. The proposed methodologies are then detailed in Sect. 5. Next Sect. 6 conducts an experimental study to demonstrate the feasibility of our proposed approaches. The future research directions are highlighted in Sect. 6. Finally, the conclusion is drawn in Sect. 6 of this chapter.

2 Background of Cyber Physical Pre-evaluation System

2.1 What is Re-manufacturability?

There is a growing interest in re-manufacturability analyses of the re-manufacturing systems since it is the key element to maintain customer satisfaction and thus company profitability. Generally speaking, re-manufacturability is the ability of used products to be easily re-manufactured and be determined by the configurable parameters, the failed state, and the re-manufacturing technology [49]. Regarding the configurable parameters, Wu [42] pointed out that the influent factors included the technological feasibility of re-manufacturing, the economic feasibility of re-manufacturing, the environmental feasibility of re-manufacturing, and the product's service ability. In a similar vein, [12] emphasized that for evaluating the re-manufacturability of used products, an integrated method in which the technology feasibility (including disassembly, cleaning, inspection and sorting, par reconditioning, machine upgrading and reassembly), economic feasibility (focusing on the re-manufacturing cost), and environmental benefits (such as energy saving, material saving and pollution reduction) should be analysed. Furthermore, some researchers (e.g., [2, 8, 51]) proposed that to enhance re-manufacturability of used products,

manufacturers should take into account the early stages of the products' designs. In the light of this statement, in [17, 38, 40], the authors stated that different design structure matrix could be used as a very useful tool to examine the relationship between the different processes in order to obtain a clear ranking of the easily activities of re-manufacturing.

2.2 Why Re-manufacturability Pre-evaluation?

In most cases, re-manufacturing processes must adopt the activity of pre-evaluation because products have not been designed to be re-manufactured [52]. This activity has extracted the "secret" affecting the success of re-manufacturing since it allows for the selective using of desired parts and/or materials. In other words, it provide a relatively efficient and effective means for a re-manufacturer to obtain feedback before the used products are admitted into the re-manufacturing plant, specially, the information about which used products/parts can be disassembled [43].

2.3 Cyber Physical Pre-evaluation System

One of the ways to evaluation such ability is through cyber-physical system (CPS), which, in our context, use sensor-embedded products with networked computing to control the evaluation processes in order to remove uncertainty to the re-manufacturing systems. An early of the successful marriage of sensor-based products and evaluation processes is radio frequency identification (RFID) tag. The advent of RFID tag is critical to automatic identification, movement tracking, access control, information collection, and evaluation of operation/system's performance. Furthermore, it is also considered by some researchers (e.g., [23–25, 29]) as one of the most technology for revolutionizing a wide range of applications including re-manufacturing and reverse logistics.

3 Problem Statement

When a used product is collected, the first step is to evaluate its re-manufacturability, which is the premise to decide whether it is worthy to re-manufacture the product. At this stage, effective and reliable systems are required to gather and evaluate product usage data. Recently, some studies (e.g., [1, 28, 53]) have reported permanent sensor embedded tagging (such as RFID) may generate valuable information for improving the efficiency of re-manufacturing process. Their analyses suggested that, since there is a high level of uncertainty about the quality of components entering the re-manufacturing process, RFID-derived

information can assist in sorting components where manual inspection is traditionally employed.

This may be true for parts that are sensitive to re-manufacture, however, one of the major puzzles that RFID classification system has posed for practitioners is the reading accuracy and system reliability after its adoption. Bearing this in mind, in this study, we are about to set up a 4-stage inspection procedure to keep the misclassification rate [53] at the lowest level. Each stage is constituted of an RFID inspecting system that is responsible for a certain type of data collection and evaluation. While the used products flow passing through these four inspection points, if any of them works improperly, the operator of the cross docking should be notified to take that certain used product out for a further inspection. Such interruption highly affects the working efficiency of a cross docking station within a re-manufacturing process and thus the reliability of the entire RFID system should be enhanced as much as possible by taking various constraints into account.

4 Mathematical Modelling

As it can be seen, our focal problem falls under the category of RFID operational level research. Nevertheless, according to our recent review [44], literature provides little guidance in addressing this issue. Therefore, in this chapter, we propose to model our focal problem as a reliability-redundancy allocation problem (RRAP).

4.1 Reliability-Redundancy Allocation Problem

The optimization of system reliability is very important in many real applications and has attracted increasing attention in academic field and a variety of engineering fields. Typically RRAP problems involve the selection of components with multiple choices and redundancy levels that produce maximum benefits, and are subject to many constraints such as the cost, weight, and volume. In general an RRAP problem can be formulated as a mixed-integer programming problem which is shown in Table 1 [41].

The goal of RRAP problem is to find the optimal combination of components and the reliabilities of the components to achieve the highest system reliability [41].

4.2 4-Stage Series System

Having the characteristics of our targeted RFID system in mind (see Sect. 2 for more details), we decide to formulate it as a 4-stage series system shown in

Table 1 General form of RRAP problem

General Form of RRAP Problem	
Maximize	
	$R_s = f(\mathbf{r}, \mathbf{n})$
Subject to	
	$g_j(\mathbf{r}, \mathbf{n}) \leq l_j, \quad j = 1, 2, \cdots, m$
where:	
R_s	the reliability of a system
$\mathbf{r} = (r_1, r_2, \cdots, r_d)$	the vector of the component reliabilities for the system
r_i	the reliability of the ith subsystem
$\mathbf{n} = (n_1, n_2, \cdots, n_d)$	the vector of redundancy allocation for the system
n_i	the number of components in the ith subsystem
$f(\cdot)$	the objective function for the overall system reliability
$g(\mathbf{r}, \mathbf{n})$	the set of constraint functions
$g_j(\mathbf{r}, \mathbf{n})$	the jth constraint function
d	the number of subsystems in the system
l	the vector of resource limitations

Eq. (1). In the literature, this formulation has also been successfully used for modelling over-speed protection system of a gas turbine [21, 50].

$$\text{Maximize} : f(\mathbf{r}, \mathbf{n}) = \prod_{i=1}^{4} \left[1 - (1 - r_i)^{n_i}\right] \tag{1}$$

Subject to:

$$g_1(\mathbf{r}, \mathbf{n}) = \sum_{i=1}^{4} v_i n_i \leq V,$$

$$g_2(\mathbf{r}, \mathbf{n}) = \sum_{i=1}^{4} \alpha_i (-T/\ln(r_i))^{\beta_i} \left[n_i + e^{0.25 n_i}\right] \leq C,$$

$$g_3(\mathbf{r}, \mathbf{n}) = \sum_{i=1}^{4} w_i n_i e^{0.25 n_i} \leq W.$$

where $0.5 \leq r_i \leq 1 - 10^{-6}$, $r_i \in R^+$, $1 \leq n_i \leq 10$, $n_i \in Z^+$, v_i represents the volume of each component in the subsystem i, V denotes the upper limit on the sum of the subsystems' products of volume, C is the upper limit on the cost of the system, $\alpha_i(-T/\ln(r_i))^{\beta_i}$ is the cost of each component with reliability r_i at subsystem i in which α_i and β_i are coefficients, T is the operating time during which the component must not fail, and W is the upper limit on the weight of the system.

5 Proposed Methodology

Many classical mathematical methods have failed to address the non-convexities and non-smoothness in RRAP problems. As an alternative to the classical optimization approaches, the CI approaches have been given much attention by many researchers because of their superior capability in finding an almost global optimal solution. In this research, we choose teaching—learning-based optimization (TLBO) and firefly algorithm (FA) as a vehicle to address our 4-stage series system problem.

5.1 Background of TLBO

Teaching—earning-based optimization (TLBO) is a new efficient population based algorithm inspired by the influence of a teacher on the output of learners in a class, which learners first acquire knowledge from a teacher (i.e., teacher phase) and then from classmates (i.e., learner phase) [35]. In principle, population consists of learners in a class and design variables are courses offered. The output in TLBO algorithm is considered in terms of results or grades of the learners which depend on the quality of teacher. That means, a high quality teacher is usually considered as a highly learned person who trains learners so that they can have better results in terms of their marks or grades. Moreover, learners also learn from the interaction among themselves which also helps in improving their results. Working of both the phase is explained below.

Teacher Phase: In the model, this phase produces a random ordered state of points called learners within the search space. Then a point is considered as the teacher, who is highly learned person and shares his or her knowledge with the learners, and others learn significant group information from the teacher. It is the first part of the algorithm where the mean of a class increases from M_A to M_B depending upon a good teacher. At this point, we assumed a good teacher is one who brings his/her learners up to his/her level in terms of knowledge. However, in practice this is not possible and a teacher can only move the mean of a class up to some extent depending on the capability of the class. This follows a random process depending on many factors [35]. Let M_i be the mean and T_i be the teacher at any iteration i. T_i will try to more mean M_i towards its own level, so now the new mean will be T_i designated as M_{new}. The solution is updated according to the difference between the existing and the new mean given by Eq. (2) [35]:

$$Difference\_Mean_i = r_i(M_{new} - T_F M_i) \tag{2}$$

The value of T_F can be either 1 or 2, which is again a heuristic step and decided randomly with equal probability as shown in Eq. (3) [35]:

$$T_F = round[1 + rand(0, 1)\{2 - 1\}]. \tag{3}$$

This difference modifies the existing solution according to Eq. (4) [35]:

$$X_{new,i} = X_{old,i} + Difference\_Mean_i \tag{4}$$

Learner Phase: It is the second part of the algorithm where learners increase their knowledge by interaction among themselves. So, a solution is randomly interacted to learn something new with other solutions in the population. In the light of this statement, a solution will learn new information if the other solutions have more knowledge than him or her. Mathematically the learning phenomenon of this phase is expressed in Eq. (5) [35]:

$$\begin{aligned} X_{new,i} &= X_{old,i} + r_i\left(X_i - X_j\right), \ \ if\ f(X_i) < f\left(X_j\right) \\ X_{new,i} &= X_{old,i} + r_i\left(X_j - X_i\right), \ \ if\ f\left(X_j\right) < f(X_i) \end{aligned} \tag{5}$$

After a number of sequential teaching–learning cycles, where the teacher convey knowledge among the learners and those level increases toward his or her own level, the distribution of the randomness within the search space becomes smaller and smaller about to point considering as teacher. It means knowledge level of the whole class shows smoothness and the algorithm converges to a solution. Also, a termination criterion can be a predetermined maximum iteration number is reached.

In many aspects, TLBO resembles evolutionary algorithms [33] such as an initial population is randomly generated; moving/learning towards teacher and classmates can be regarded as a special mutation operator; and selection is deterministic (i.e., two solutions are compared and the better one always survives) [11]. The TLBO algorithm has been used in solving many problems, remarkable results have been reported about TLBO outperforming many algorithms such as differential evolution [39], evolutionary strategies [6], and particle swarm optimization [26].

5.2 Background of FA

The firefly algorithm (FA) is a meta-heuristic, nature-inspired, optimization algorithm which is based on the social (flashing) behaviour of fireflies, or lighting bugs, in the summer sky in the tropical temperature regions [45–47]. It was developed by Dr. Xin-She Yang at Cambridge University in 2007, and it is based on the swarm behaviour such as fish, insects, or bird schooling in nature. In the FA, physical entities (fireflies) are randomly distributed in the search space. They carry a bio-luminescence quality, called luciferin, as a signal to communicate with other fireflies, especially to prey attractions [5]. In detail, each firefly is attracted by the brighter glow of other neighbouring fireflies. The attractiveness decreases as their distance increases. If there is no brighter one than a particular firefly, it will move randomly. Its main advantage is the fact that it uses mainly real random numbers,

and it is based on the global communication among the swarming particles (i.e., the fireflies), and as a result, it seems more effective in multi-objective optimization. Normally, FA uses the following three idealized rules [47] to simplify its search process to achieve an optimal solution:

- Fireflies are unisex so that one firefly will be attracted to other fireflies regardless of their sex, that means no mutation operation will be done to alter the attractiveness fireflies have for each other;
- The sharing of information or food between the fireflies is proportional to the attractiveness that increases with a decreasing Cartesian or Euclidean distance between them due to the fact that the air absorbs light. Thus for any two flashing fireflies, the less bright one will move towards the brighter one. If there is no brighter one than a particular firefly, it will move randomly;
- The brightness of a firefly is determined by the landscape of the objective function. For the maximization problems, the light intensity is proportional to the value of the objective function.

Furthermore, there are two important issues in the FA that are the variation of light intensity or brightness and formulation of attractiveness. Yang [45] simplifies the attractiveness β of a firefly is determined by its brightness I which in turn is associated with the encoded objective function. As light intensity and thus attractiveness decreases as their distance from the source increases, the variations of light intensity and attractiveness should be monotonically decreasing functions.

- Variation of Light Intensity: Suppose that there exists a swarm of n fireflies, and x_i, $i = 1, 2, \ldots, n$ represents a solution for a firefly i initially positioned randomly in the space, whereas $f(x_i)$ denotes its fitness value. In the simplest form, the light intensity $I(r)$ varies with the distance r monotonically and exponentially. That is determined through Eq. (6) [45–47]:

$$I = I_0 e^{-\gamma r_{ij}} \tag{6}$$

where I_0 is the original light intensity, γ is the light absorption coefficient, and r is the distance between firefly i and firefly j at x_i and x_j as Cartesian distance $r_{ij} = \left\|x_i - x_j\right\| = \sqrt{\sum_{k=1}^{d} \left(x_{i,k} - x_{j,k}\right)^2}$ or the ℓ_2 - norm, where $x_{i,k}$ is the kth component of the spatial coordinate x_i of the ith firefly and d is the number of dimensions we have, for $d = 2$, we have $r_{ij} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2}$.

- Movement toward attractive Firefly: A firefly attractiveness is proportional to the light intensity seen by adjacent fireflies [45]. Each firefly has its distinctive attractiveness β which implies how strong it attracts other members of the swarm. However, the attractiveness is relative; it will vary with the distance between two fireflies. The attractiveness function $\beta(r)$ of the firefly is determined via Eq. (7) [45–47]:

$$\beta = \beta_0 e^{-\gamma r_{ij}^2} \tag{7}$$

where, β_0 is the attractiveness at $r = 0$, and γ is the light absorption coefficient which controls the decrease of the light intensity.

- The movement of a firefly i at location x_i attracted to another more attractive (brighter) firefly j at location x_j is determined based on Eq. (8) [45–47]:

$$x_i(t+1) = x_i(t) + \beta_0 e^{-\gamma r_{ij}^2}\left(x_j - x_i\right) + \alpha\varepsilon_i \tag{8}$$

where, the first term is the current position of a firefly, the second term is used for considering a firefly's attractiveness to light intensity seen by adjacent fireflies, and the third term is randomization with the vector of random variables ε_i being drawn from a Gaussian distribution, in case there are not any brighter ones. The coefficient α is a randomization parameter determined by the problem of interest.

- Special Cases: From Eq. (8), it is easy to see that there exit two limit cases when γ is small or large, respectively [45–47]. When γ tends to zero, the attractiveness and brightness are constant $\beta = \beta_0$ which means the light intensity does not decrease as the distance r between two fireflies increases. Therefore, a firefly can be seen by all other fireflies, a single local or global optimum can be easily reached. This limiting case corresponds to the standard particle swarm optimization algorithm. On the other hand, when γ is very large, then the attractiveness (and thus brightness) decreases dramatically, and all fireflies are short-sighted or equivalently fly in a deep foggy sky. This means that all fireflies move almost randomly, which corresponds to a random search technique.

In general, the FA corresponds to the situation between these two limit cases, and it is thus possible to fine-tune these parameters, so that FA can find the global optima as well as all the local optima simultaneously in a very effective manner. A further advantage of FA is that different fireflies will work almost independently, it is thus particular suitable for parallel implementation. It is even better than genetic algorithm and particle swarm optimization because fireflies aggregate more closely around each optimum. It can be anticipated that the interactions between different sub-regions are minimal in parallel implementation. Nowadays, mechanisms of firefly communication via luminescent flashes and their synchronization has been used effectively to solve the problems in various areas, such as in continuous constrained optimization [32], economic emissions load dispatch [3], image compression [18, 20], mixed variable structural optimisation [15], re-machining parameter optimization [43], scheduling [36], clustering [19, 37], parameter tuning [48], wireless network design [31], dynamic marketing pricing [22].

Table 2 Benchmark test function (adapted from [35])

Benchmark test function
Minimize:
$f(\mathbf{x}) = (x_1-10)^2+5(x_2-12)^2+x_3^4+3(x_4-11)^2+10x_5^6+7x_6^2+x_7^4-4x_6x_7-10x_6-8x_7$
Subject to:
$g_1(\mathbf{x}) = -127+2x_1^2+3x_2^4+x_3+4x_4^2+5x_5 \le 0$
$g_2(\mathbf{x}) = -282+7x_1+3x_2+10x_3^2+x_4-x_5 \le 0$
$g_3(\mathbf{x}) = -196+23x_1+x_2^2+6x_6^2-8x_7 \le 0$
$g_4(\mathbf{x}) = 4x_1^2+x_2^2-3x_1x_2+2x_3^2+5x_6-11x_7 \le 0$
where:
$-10 \le x_i \le 10\,(i=1,2,\ldots 7)$

Table 3 Comparison of results (10 runs) obtained by using FA and TLBO for Benchmark test function

	Benchmark test function	
	FA	TLBO
$f(\mathbf{x}^*)$—best	680.7249	680.6305
$f(\mathbf{x}^*)$—worst	681.1270	680.6325
$f(\mathbf{x}^*)$—mean	680.8900	680.6318
$\mathbf{x}^*$—best	$\begin{pmatrix} 2.349317 \\ 1.948980 \\ -0.442630 \\ 4.366729 \\ -0.638468 \\ 1.027457 \\ 1.614927 \end{pmatrix}$	$\begin{pmatrix} 2.331588 \\ 1.951348 \\ -0.477926 \\ 4.365752 \\ -0.626203 \\ 1.032656 \\ 1.593086 \end{pmatrix}$

$f(\mathbf{x}^*)$ Objective Function Value; $\mathbf{x}^*$ Optimum Solution

5.3 Benchmark Test Function

In this section, a benchmark test function (see Table 2) is selected from the literature to demonstrate the effectiveness of proposed TLBO and FA in dealing with function optimization problem. The chosen function is a non-linear minimization problem which has seven design variables and four non-linear inequality constraints.

The parameter settings for TLBO and FA are as follows:

- TLBO: Population size is 50, generations are 2,000, total number of function evaluations are 100,000;
- FA: Population size is 20, generations are 5,000, total number of function evaluations are 100,000.

The numerical results obtained via FA and TLBO are outlined in Table 3.

Table 4 Benchmark engineering design problem (adapted from [35])

Benchmark engineering design problem	
Minimize	$f(x) = 0.6224d_1rL + 1.7781d_2r^2 + 3.1661b_1^2L + 19.84d_1^2r$
Subject to	$g_1(\mathbf{x}) = -d_1 + 0.0193r \le 0$
	$g_2(\mathbf{x}) = -d_2 + 0.00954r \le 0$
	$g_3(\mathbf{x}) = -\pi r^2 L - \frac{4}{3}\pi r^3 + 1296000 \le 0$
	$g_4(\mathbf{x}) = L - 240 \le 0$
where	$0 \le d_1 \le 99 \quad 0 \le d_2 \le 99$
	$10 \le r \le 200 \quad 10 \le L \le 200$

The optimum solution is at $x^* = (2.330499, 1.951372, -0.4775414, 4.365726, -0.6244870, 1.1038131, 1.594227)$ with objective function value $f(x^*) = 680.6300573$ [35]. Although, we only test ten runs on each method, it can be observed from Table 3, both FA and TLBO work fine on solving benchmark test function.

5.4 Benchmark Engineering Design Optimization Problem

In this section, a constrained benchmark engineering design problem (i.e., pressure vessels design) is selected from the literature to test the effectiveness of proposed TLBO and FA in dealing with constrained optimization problem. For a given volume and working pressure, the basic aim of designing a pressure vessel is to get the total cost minimized. The typical design variables are such as the thickness of the head and body, the inner radius, and the length of the cylindrical section. This is a well-known test problem for optimization and the standard form can be found in Table 4.

To evaluate the performance of FA and TLBO algorithms for optimizing the design of pressure vessels, we also run the simulation 10 times under the same parameter settings as previously mentioned. The numerical results are shown in Table 5.

As it shown in Table 5, according to other similar studies such as [9], both solutions obtained through TLBO and FA are reasonable in which TLBO can generate a slightly better results than FA on the test problem.

From the above mentioned two examples, we can see that TLBO and FA are very good optimizer and suitable for many applications.

Table 5 Comparison of results (10 runs) obtained by using FA and TLBO for Benchmark engineering design problem

	Benchmark engineering design problem	
	FA	TLBO
$f(\mathbf{x}^*)$—best	6117.2432	5885.3330
$f(\mathbf{x}^*)$—worst	6418.8358	5885.3890
$f(\mathbf{x}^*)$—mean	6263.9090	5885.3453
$\mathbf{x}^*$—Best	$\begin{pmatrix} 0.828478 \\ 0.425984 \\ 42.619529 \\ 172.301498 \end{pmatrix}$	$\begin{pmatrix} 0.778169 \\ 0.384649 \\ 40.319619 \\ 200 \end{pmatrix}$

$f(\mathbf{x}^*)$ Objective Function Value; $\mathbf{x}^*$ Optimum Solution

Table 6 Data used for RFID system

Stage	$10^5 A_I$	B_I	V_I	W_I	V	C	W	T
1	1.0	1.5	1	6	200	500	450	2000
2	2.3	1.5	2	6	–	–	–	–
3	0.3	1.5	3	8	–	–	–	–
4	2.3	1.5	2	7	–	–	–	–

Table 7 Convergence results (10 runs) obtained by using FA and TLBO

	Convergence results of f(r, n)	
	FA	TLBO
$f(\mathbf{r}, \mathbf{n})$—best	99.6152	99.6156
$f(\mathbf{r}, \mathbf{n})$—worst	99.6118	99.6153
$f(\mathbf{r}, \mathbf{n})$—mean	99.6139	99.6155

6 Experimental Study by Using TLBO and FA

Witnessing the capability of TLBO and FA, in this section, we decide to carry out an experimental study to solve our focal problem. Suppose that we have the following parameters and constraints (arrayed in Table 6) for our RFID system.

Based on Eq. (1), we run FA and TLBO 10 times for each and the simulation results are listed in Tables 7 and 8, respectively.

From Table 8, we can see that in order to keep our RFID system reliability at the highest level, the components number and the corresponding reliability should be designed based on the results obtained via TLBO and FA.

Table 8 Best results (10 runs) obtained by using FA and TLBO for RFID system

	RFID system	
	FA	TLBO
$f(\mathbf{R}, \mathbf{N})$	99.6152	99.6156
N_1	5	5
n_2	5	5
n_3	4	4
n_4	5	5
r_1	72.0518	72.4646
r_2	74.0182	73.9552
r_3	82.1285	81.9053
r_4	74.8874	74.7276

7 Future Research Directions

Since our targeted question belongs to a class of constrained nonlinear mixed-integer programming problem which means the solution of this kind of problems consists two parts, i.e., a real part and an integer part. As the searching space and complexity of these two parts are different, it might be more promising to use different searching mechanisms to obtain individual optimal solution for each of these two parts. Therefore one possible future research direction is to employ two algorithms to search the real part and the integer part, respectively.

In order to maximize a system's reliability, except the reliability-redundancy allocation solution, one can also consider other options such as enhancing the component reliability [30]. Since there are many other communication systems may utilize the similar frequencies within the communication range of RFID which in turn could interfere the reliability of the RFID system, a good filter design for RFID receiver is always a necessary. Therefore an immediate extension of the current research would be employing suitable CI methods to optimize the performance of signal filter component within a RFID reader.

8 Conclusions

Re-manufacturability classification based on radio frequency identification (RFID) system is a great concept transition and innovation. The idea is to "take the initiative to prevent problems", which can greatly save resources and energy of the whole world and bring enormous economic benefits as well as social benefits. However, recently the growing interest in cyber physical re-manufacturability pre-evaluation faces major challenges due to the error prone nature of RFID devices. The focus of our work is complementary to the inherent unreliability of RFID systems, and ask whether the reliability can be improved using more redundant components (i.e., RFID readers) in wide range type series. In this chapter, we first

formulate our focal scenario as a reliability-redundancy allocation problem (RRAP). Then, two of the recently developed computational intelligence approaches called teaching—learning-based optimization (TLBO) which is based on the effect of the influence of a teacher on the output of learners in a class, and firefly algorithm (FA) which is based on the social (flashing) behaviour of fireflies, or lighting bugs, in the summer sky in the tropical temperature regions, are employed to address our focal problem. Simulation results suggest that the proposed TLBO and FA are viable optimization techniques in improving the RFID classification system's reliability.

References

1. M. AliIlgin, S.M. Gupta, Performance improvement potential of sensor embedded products in environmental supply chains. Resour. Conserv. Recycl. **55**(6), 580–592 (2011).
2. T. Amezquita, R. Hammond, M. Salazar, B. Bras, Characterizing the re-manufacturability of engineering systems. Paper presented at the ASME advances in design automation conference, Boston, Massachusetts, USA, pp. 271–278, 17–20 Sept 1995
3. T. Apostolopoulos, A. Vlachos, Application of the firefly algorithm for solving the economic emissions load dispatch problem. Int. J.Combinatorics, Article ID **523806**, 1–23 (2011)
4. N. Aras, T. Boyaci, V. Verter, The effect of categorizing returned products in re-manufacturing. IIE Trans. **36**, 319–331 (2004)
5. B.G. Babu, M. Kannan, Lightning bugs. Resonance **7**(9), 49–55 (2002)
6. H.-G. Beyer, H.-P. Schwefel, Evolution strategies: a comprehensive introduction. J. Nat Comput. **1**(1), 3–52 (2002)
7. J.D. Blackburn, V.D.R. Guide, G.C. Souza, L.N.V. Wassenhove, Reverse supply chains for commercial returns. Calif. Manag. Rev. **46**(2), 6–22 (2004)
8. B. Bras, M.W. McIntosh, Product, process, and organizational design for re-manufacture—an overview of research. Rob. Comput. Integr. Manuf. **15**, 167–178 (1999)
9. L.C. Cagnina, S.C. Esquivel, C.A. Coello, Solving engineering optimization problems with the simple constrained particle swarm optimizer. Informatica **32**, 319–326 (2008)
10. D.J. Cook, J.C. Augusto, V.R. Jakkula, Ambient intelligence: technologies, applications, and opportunities. Pervasive Mob. Comput. **5**, 277–298 (2009)
11. M. Črepinšek, S.-H. Liu, L. Mernik, A note on teaching–learning-based optimization algorithm. Inf. Sci. **212**, 79–93 (2012)
12. Y. Du, H. Cao, F. Liu, C. Li, X. Chen, An integrated method for evaluating the re-manufacturability of used machine tool. J. Clean. Prod. **20**, 82–91 (2012)
13. M. Fleischmann, H.R. Krikke, R. Dekker, S.D.P. Flapper, A characterisation of logistics networks for product recovery. OMEGA **28**, 653–666 (2000)
14. M. Fleischmann, JAEEv Nunen, B. Gräve, Integrating closed-loop supply chains and spare-parts management at IBM. Interfaces **33**(6), 44–56 (2003)
15. A.H. Gandomi, X.-S. Yang, A.H. Alavi, Mixed variable structural optimization using firefly algorithm. Comput. Struct. **89**, 2325–2336 (2011)
16. V.D.R. Guide, R.H. Teunter, L.N.V. Wassenhove, Matching demand supply to maximize profits from re-manufacturing. Manuf. Serv. Oper. Manage. **5**(4), 303–316 (2003)
17. R. Hammond, T. Amezquita, B. Bras, Issues in the automotive parts re-manufacturing industry-a discussion of results from surveys performed among re-manufacturers. Int. J. Eng. Des. Autom.—Spec Issue Environmentally Conscious Des. Manufact. **4**(1), 27–46 (1998)
18. M.-H. Horng, Vector quantization using the firefly algorithm for image compression. Expert. Syst. Appl. **39**, 1078–1091 (2012)

19. M.-H. Horng, Y.-X. Lee, M.-C. Lee, R.-J. Liou, *Firefly meta-heuristic algorithm for training the radia basis function network for data classification and disease diagnosis*. in ed. by R. Parpinelli Theory and New Applications of Swarm Intelligence, Chapter 7, pp. 115–132: In-Tech (2012)
20. M.-H. Horng, R.-J. Liou, Multilevel minimum cross entropy threshold selection based on the firefly algorithm. Expert. Syst. Appl. **38**, 14805–14811 (2011)
21. Y.-C. Hsieh, P.-S. You, An effective immune based two-phase approach for the optimal reliability. Appl. Math. Comput. **218**, 1297–1307 (2011)
22. J. Jumadinova, P. Dasgupta, Firefly-inspired synchronization for improved dynamic pricing in online markets. Paper presented at the 2nd IEEE international conference on self-adaptive and self-organizing Systems, pp. 402–412, (2008)
23. Ö. Karaer, H.L. Lee, Managing the reverse channel with RFID-enabled negative demand information. Prod. Oper. Manage. **16**(5), 625–645 (2007)
24. M. Kärkkäinen, T. Ala-Risku, K. Främling, The product centric approach: a solution to supply network information management problems? Comput. Ind. **52**(2), 147–159 (2003)
25. M. Kärkkäinen, T. Ala-Risku, J. Holmström, Increasing customer value and decreasing distribution costs with merge-in-transit. Int. J. Phys. Distrib. Logistics. Manage. **33**(2), 132–148 (2003)
26. J. Kennedy, R. C. Eberhart, Particle swarm optimization. Paper presented at the IEEE International Joint conference on neural networks (1995)
27. H. Krikke, I. l. Blanc, S. v. d. Velde, Product modularity and the design of closed-loop supply chains. Calif. Manage. Rev. **46**(2), 23–38 (2004)
28. A. Kulkarni, D. Ralph, D. McFarlane, Value of RFID in re-manufacturing. Int. J. Serv. Oper. Inf. **2**(3), 225–252 (2007)
29. A. G. Kulkarni, A. K. N. Parlikad, D. C. McFarlane, M. Harrison, Networked RFID systems in product recovery management. Paper presented at the IEEE international symposium on electronics and the environment (ISEE 2005), pp. 66–71 (2005)
30. W. Kuo, V.R. Prasad, An annotated overview of system-reliability optimization. IEEE Trans. Reliab. **49**, 176–187 (2000)
31. R. Leidenfrost, W. Elmenreich, Establishing wireless time-triggered communication using firefly clock synchronization approach. Paper presented at the 2008 international workshop on intelligent solutions in embedded systems, pp. 1–8, (2008)
32. S. Łukasik, S. Żak, *Firefly algorithm for continuous constrained optimization tasks Computational Collective Intelligence.* Semantic Web, Social Networks and Multiagent Systems, LNCS 5796. (Berlin, Spinger-Verlag, 2009), pp. 97–106
33. Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Programs*, 3rd edn. (Springer-Verlag, Berlin Heidelberg, 1996)
34. J. A. E. E. v. Nunen, R. Zuidwijk, E-enabled closed-loop supply chains. Calif. Manage. Rev. **46**(2), 40–54, (2004)
35. R.V. Rao, V.J. Savsani, D.P. Vakharia, Teaching–learning-based optimization: a novel method for constrained mechanical design optimization problems. Comput. Aided. Des. **43**, 303–315 (2011)
36. M.K. Sayadi, R. Ramezanian, N. Ghaffari-Nasab, A discrete firefly meta-heuristic with local search for makespan minimization in permutation flow shop scheduling problems. Int. J. Ind. Eng. Comput. **1**, 1–10 (2010)
37. J. Senthilnath, S.N. Omkar, V. Mani, Clustering using firefly algorithm: performance study. Swarm Evol. Comput. **1**, 164–171 (2011)
38. L. Shu, W. Flowers, A structured approach to design for re-manufacture. Intell. Concurrent Des.: Fundam. Methodol. Model. Pract. **66**, 13–19 (1993)
39. R. Storn, K. Price, Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. J. Global. Optim. **11**(4), 341–359 (1997)
40. E. Sundin, B. Bras, Making functional sales environmentally and economically beneficial through product re-manufacturing. J. Clean. Prod. **13**, 913–925 (2005)

41. L. Wang, L.-P. Li, A coevolutionary differential evolution with harmony search for reliability–redundancy optimization. Expert. Syst. Appl. **39**, 5271–5278 (2012)
42. X. Wu, Research on design management based on green re-manufacturing engineering. Syst. Eng. Procedia **4**, 448–454 (2012)
43. B. Xing, W.-J. Gao, Computational intelligence in re-manufacturing. 701 E. Chocolate Avenue, Suite 200, Hershey PA 17033: IGI Global, ISBN 978-1-4666-4908-8 (2014)
44. B. Xing, W.-J. Gao, T. Marwala, The applications of computational intelligence in radio frequency identification research. Paper presented at the IEEE international conference on systems, man, and cybernetics (IEEE SMC), 14–17 Oct, (Seoul, Korea, 2012) pp. 2067–2072
45. X.-S. Yang, *Nature-inspired metaheuristic algorithms* 2nd edn. (UK, Luniver Press) ISBN 978-1-905986-28-6 (2008)
46. X.-S. Yang, Firefly algorithms for multimodal optimization, in *SAGA 2009, LNCS 5792*, ed. by O. Watanabe, T. Zeugmann (Springer-Verlag, Berlin Heidelberg, 2009), pp. 169–178
47. X.-S. Yang, Firefly algorithm, stochastic test functions and design optimisation. Int. J. Bio-Inspired Comput. **2**(2), 78–84 (2010)
48. X.-S. Yang, Chaos-enhanced firefly algorithm with automatic parameter tuning. Int. J. Swarm Intell. Res. **2**(4), 1–11 (2011)
49. J. Yao, S. Zhu, The research of design system for re-manufacturing. N. Technol. N. Process **5**, 22–24 (2004)
50. W.-C. Yeh, T.-J. Hsieh, Solving reliability redundancy allocation problems using an artificial bee colony algorithm. Comput. Oper. Res. **38**, 1465–1473 (2011)
51. H. Yüksel, Design of automobile engines for re-manufacture with quality function deployment. Int. J. Sustain. Eng. **3**(3), 170–180 (2010)
52. Z.-N. Zhang, Z.-L. Liu, Y. Chen, Y.-B. Xie, Knowledge flow in engineering design: an ontological framework. Proc. Ins. Mech. Eng. Part C: J. Mech. Eng. Sci. **227**(4), 760–770 (2013)
53. C. Zikopoulos, G. Tagaras, On the attractiveness of sorting before disassembly in re-manufacturing. IIE Trans. **40**, 313–323 (2008)